NASA CONTRACTOR REPORT

NASA CR-453

INFORMATION REQUIREMENTS FOR GUIDANCE AND CONTROL SYSTEMS

by John Peschon, Lewis Meier III, Robert E. Larson, Wade H. Foy, Jr., and Charles H. Dawson

Prepared under Contract No. NAS 2-2457 *by*
STANFORD RESEARCH INSTITUTE
Menlo Park, Calif.
for Ames Research Center

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION - WASHINGTON, D. C. - MAY 1966

NASA CR-453

INFORMATION REQUIREMENTS FOR GUIDANCE AND CONTROL SYSTEMS

By John Peschon, Lewis Meier III, Robert E. Larson, Wade H. Foy, Jr., and Charles H. Dawson

Prepared under Contract No. NAS 2-2457 by
STANFORD RESEARCH INSTITUTE
Menlo Park, Calif.

for Ames Research Center

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION

For sale by the Clearinghouse for Federal Scientific and Technical Information
Springfield, Virginia 22151 – Price $5.00

ABSTRACT

The results of research performed at SRI for NASA Ames Research Center on Contract NAS 2-2457 are summarized. The subject of this research is the dependence of the performance of a control or guidance system upon the information-handling characteristics of such key components as sensors. The disciplines of information theory and control theory are used to consider both the quantity and value of information.

CONTENTS

CONTENTS

ILLUSTRATIONS

TABLE

SUMMARY

25547

The present final report summarizes the work accomplished from November 1964 to September 1965. Its main theme is the dependence of the performance of a control or guidance system upon the information-handling characteristics of its key constituents, notably the measurement and control subsystems.

To determine this dependence, the disciplines of control theory and information theory are reviewed and extended. The notions of quantity of information and value of information are shown to be the principal measures of information; however, entropy (the measure of quantity) is of very limited assistance for the design of control and guidance systems.

The dependence of system performance on the information-handling characteristics of its constituents is investigated for optimum controllers and fixed controllers. The case of optimum controllers is resolved by the newly developed theory of combined optimization, where the optimum control decision is determined by the prior information and the measurements received. The case of fixed controllers is resolved by consideration of stochastic difference equations as well as known methods of classical control theory. The sensitivity of performance with respect to the relevant design parameters provided in both cases is particularly useful for system design and evaluation.

Additional results relating to the optimum design of quantized control systems, information systems, and adaptive systems are given.

Author

I INTRODUCTION

This report summarizes the results of a one-year study performed by Stanford Research Institute for the NASA Ames Research Center under Contract NAS 2-2457. The main problem pursued during the course of the study, namely the combined optimization problem, was formulated in an interim technical report[1*] and a solution, supported by examples, is contained in a NASA Report.[2]

The present volume reports on the objectives of the study, summarizes and relates the main findings, and justifies the approach taken.

A. OBJECTIVES OF THE STUDY

The main objective of the study is to "relate the performance of control and guidance systems to the characteristics of the information links between the various constituents, with the aim of providing exact and approximate synthesis techniques for an efficient design."

Specifically, answers to the following questions are desired:

- Of which states and parameters should measurements be made while the system is operating?
- How good must these measurements be in order to permit a stated and specified measure of system performance to be achieved?
- What are the trade-offs between alternative measurement systems, *i.e.*, different grades or types of sensors?
- How can *a priori* knowledge about the system obviate the need for acquiring information by means of measurements?
- When is it possible and justifiable to achieve stated system performance as a result of elaborate information processing as opposed to accurate sensing?

---

* References are listed at the end of the report.

- How does one derive optimum control decisions, given a set of measurements which are ordinarily incomplete and corrupted by noise?

From these objectives, it is clear that the concept of *information* plays a central role and that quantitative descriptions of information must be defined to answer the questions previously raised. Since control system theory and information theory are the two scientific disciplines most closely related to the central theme of the study, their potential was to be critically reviewed and, if warranted, extended. One important question in need of a precise answer was "What is the usefulness of such information-theoretic measures as entropy to aid in the synthesis of control and guidance systems?"

B. JUSTIFICATION FOR THE STUDY

With the advent of complex systems, such as used for space navigation, exploration, and communication, the dollar cost of the sensors and the information links has become a significant part of the total system cost. Synthesis techniques allowing the specified systems objective to be accomplished with the least costly combination of sensors and communication links consequently have an obvious economic justification. Furthermore, the forthcoming generation of relatively inexpensive and compact digital computers (integrated circuits) in many cases make it possible, by elaborate data processing, to obviate the need for very accurate sensing. Finally, in many advanced missions, the objective can be satisfied only if the operation of the various subsystems is optimized; consequently, a very high premium is placed upon the complete utilization of all information to generate optimum decisions.

As regards the scientific justifications, it is noted that an ever increasing fraction of the engineering, mathematical, military, and business communities are concerned with the problem of generating decisions (or delegating this function to computers) based on the information available about the systems that these decisions affect. This preoccupation has given rise to several scientific disciplines—such as control theory, information theory, operations research, theory of games and others—that may will soon become unified under the general heading of Information Sciences.

In all of these disciplines, the information and decision processes (as opposed, for example, to energy considerations) play a central role. It is consequently of considerable scientific interest to understand what quantitative descriptions of information can be defined, and how the information and decision processes are related among themselves as well as to the objectives pursued by these systems.

Although the remainder of this report is concerned with "physical" systems, such as found in control and guidance applications, it is emphasized that the general conclusions as well as the specific mathematical developments are applicable directly to "non-physical" systems, such as found in operations research, strategy, and econometrics.

C. TERMINOLOGY AND NOTATION

It is assumed that a dynamic process, also referred to as the *plant* or *signal-generating process*,

$$x_{k+1} = f(x_k, u_k, w_k, k) \qquad \text{(I-1)}$$

is given. The *state* x_k is a non-unique set of numbers that, together with the known parameters, completely describes the condition of the dynamic process at the discrete time k. The decision or control vector u_k allows the modification of state x_{k+1} in some desirable way. The disturbance or perturbation vector w_k alters the state in an arbitrary and usually unpredictable way. The time variable accounts for the known and predictable effects and parameter changes.

In the remainder of this report, difference equations will usually be preferred to differential equations for two reasons, namely

- The computation of the control laws (derived in later sections) generally requires a digital computer.
- Any whiteness assumption made about such random effects as w_k will be much more meaningful and justifiable in discrete time than in continuous time.

To proceed, a categorization of the main classes of problems pursued in control theory during the recent past is given.

1. The Optimum Control Problem

The optimum control problem (Fig. I-1), for which many practical solutions are available, is defined as follows:

Given the present state x_k and the process description (I-1) with $w_i \equiv 0 (i = k, \ldots, N)$.

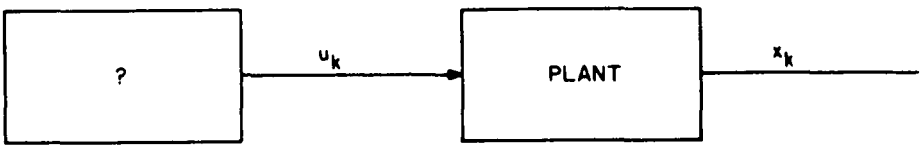

FIG. I-1 THE OPTIMUM CONTROL PROBLEM

Find a sequence of controls $u_k, \ldots, u_{N-1}$ that minimizes a variational cost expression of the form

$$J(x_k; u_k, \ldots, u_N) = \sum_{i=k}^{N} l(x_i, u_i, i) \qquad \text{(I-2)}$$

subject to the constraints

$$u_i \in U$$

$$x_i \in X \quad . \qquad \text{(I-3)}$$

Note that only the state x_k need be known to determine the whole sequence $u_k, \ldots, u_N$, since future states x_i are completely predictable. In other words

$$u_i = u_i(x_k) \qquad i \geq k \quad .$$

2. The Optimum Stochastic Control Problem

a. STANDARD FORMULATION

The optimum stochastic control problem (Fig. I-2), for which solutions based on dynamic programming are available,[3] is defined as follows,

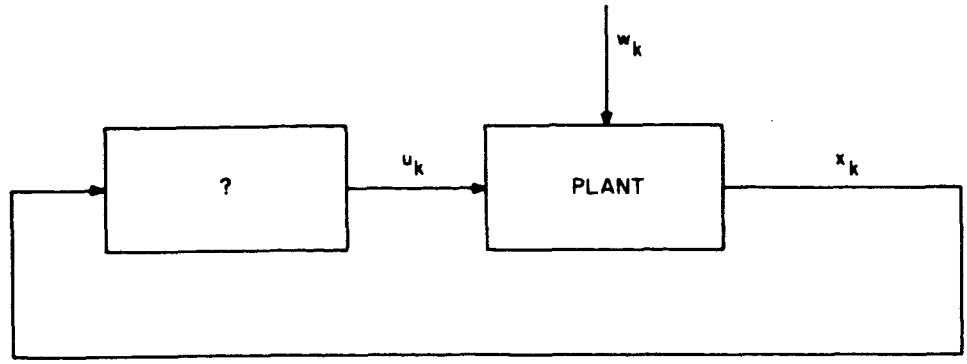

FIG. I-2 THE OPTIMUM STOCHASTIC CONTROL PROBLEM

Given the present state x_k, the process description (I-1) with $w_i \neq 0 (i = k, \ldots, N)$, and the probability density function of w_i, $p(w_i)$

Find the sequence of controls

$$u_i = u_i(x_i) \tag{I-4}$$

that on the average minimizes a variational cost expression of the form

$$J(x_k; u_k, \ldots, u_N) = \underset{i=k, \ldots, N}{E} \left\{ \sum_{i=k}^{N} l(x_i, u_i, i) \right\} \tag{I-5}$$

Note that at every stage k, the actual state x_k resulting from the application of u_{k-1} and w_{k-1} is assumed to be known exactly from appropriate measurements.

b. NONSTANDARD FORMULATION

It is entirely possible, but not customary, to formulate an optimum stochastic control problem where no measurements x_k are made and where consequently the optimum control sequence [which minimizes (I-5)] is made to depend on the initial state x_0 only

$$u_i = u_i(x_0)$$

This may be viewed as an open-loop solution, whereas the standard formulation leads to a feedback solution. The justification for

making measurements is determined by comparison of the minimum costs J in both cases.

3. The Optimum Estimation Problem

The optimum estimation problem (Fig. I-3), to which a very general solution has become available quite recently,[4] is formulated as follows:

Given a signal-generating process of the form (I-1), a measurement process described by

$$z_k = h(x_k, v_k, k) \tag{I-6}$$

where v_k is the measurement noise, and the probability density functions $p(x_0)$, $p(w_i)$ and $p(v_i)$, $i = 0, \ldots, N$

Find the conditional probability density function of the state

$$p(x_k / z_0, \ldots, z_k) \quad . \tag{I-7}$$

From (I-7), the conditional moments, such as the conditional mean denoted by $\hat{x}_{k/k}$, are derived easily.

A special case of the general optimum estimation problem is the well-known Kalman-Bucy filter.

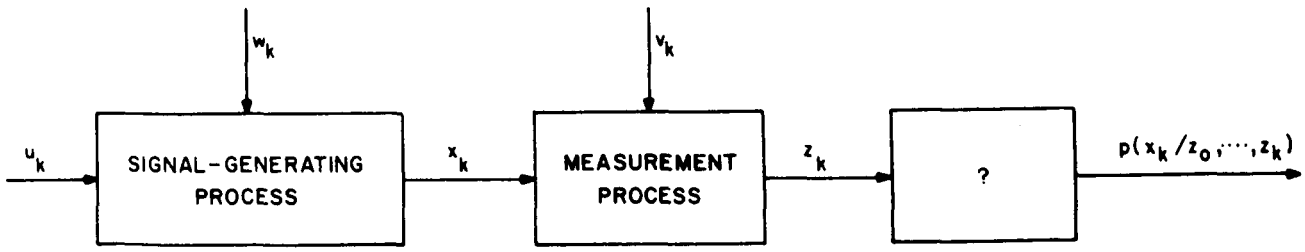

FIG. I-3 THE OPTIMUM ESTIMATION PROBLEM

4. The Combined Optimization Problem

The combined optimization problem (Fig. I-4), for which a solution was found in the course of this contract,[2] combines the optimum stochastic control problem with the optimum estimation problem in the general case.

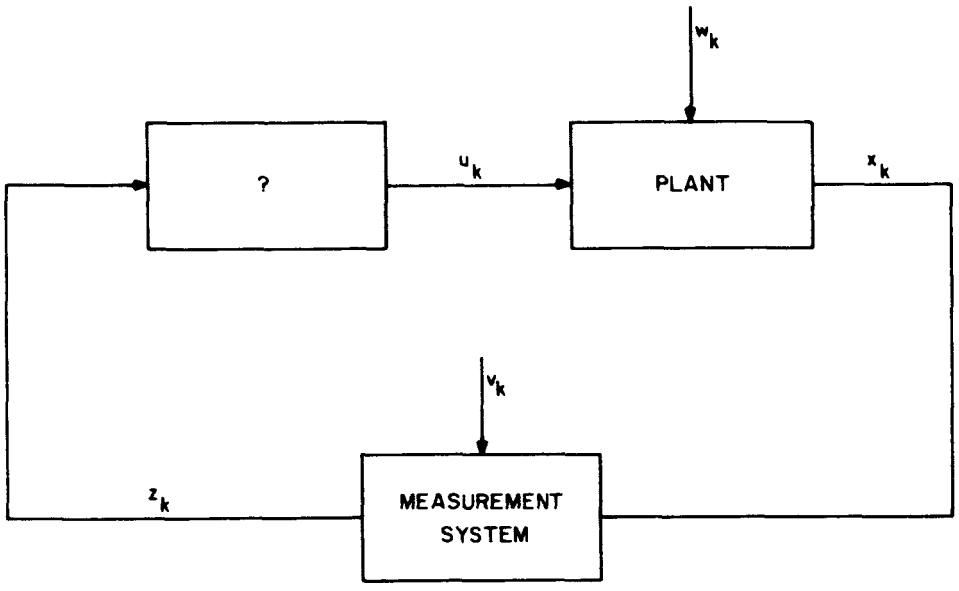

FIG. I-4 THE COMBINED OPTIMIZATION PROBLEM

It is defined as follows:

Given the state Eq. (I-1), the measurement Eq. (I-6), and the probability density functions $p(x_0)$, $p(w_i)$, $p(v_i)$ $(i = 0,\ldots,N)$,

Find the sequence of controls u_k that on the average minimizes a variational cost expression of the form

$$J(z_0,\ldots,z_k;u_k,\ldots,u_N) = \underset{y}{E}\left\{\sum_{i=k}^{N} l(x_i,u_i,i)\right\} \qquad \text{(I-8)}$$

where the vector y encompasses all the random variables, *i.e.*, x_0, w_i and v_i. The optimum control u_k to be applied at the present time k depends on all the past measurements

$$u_k = u_k(z_0,\ldots,z_k) \quad . \qquad \text{(I-9)}$$

In Sec. II, it is shown that the general combination of plant and sensing equipment can be modeled by the state Eq. (I-1) and the measurement Eq. (I-6).

Much of this report will be devoted to the combined optimization problem, which answers the question "What is the best performance obtainable with the noisy and generally incomplete measurements z_t" and hence relates the system's performance to the characteristics of the measurement system under the assumption that the measurements z_t are utilized optimally.

The combined optimization problem is easily extended to the adaptive case where the plant (I-1) is not accurately known before measurements z have been received. It thus provides a rigorous and much needed mathematical framework for adaptive systems and other recently evolved system concepts, as will be discussed in detail in Sec. III.

A real shortcoming of the combined optimization problem solution obtained, however, is the impossibility of computing (with present-day machines) an exact solution in the general case. Important special cases, fortunately, are amenable to machine computation, and fairly evident approximations can be used in the general case.

5. Additional Problem Formulations

At this time (1965), several important additional problems are being formulated in the field of control theory, notably the theory of differential games and the theory of optimum classification. Both are strongly related to the formulations discussed above in the sense that similar mathematical approaches apply. These additional formulations will not be treated further in any detail, despite their strong relation to the subject matter of this report.

6. Information-Theoretic Concepts

In what follows, the information-theoretic concept of entropy (or uncertainty) will sometimes be used.

Given a probability density function $p(y)$, where y is an n-dimensional vector in space and/or time, the entropy is defined as

$$H = -\int_y p(y) \log p(y) dy \quad . \tag{I-10}$$

For a Gaussian distribution

$$p(y) = \frac{1}{(\sqrt{2\pi})^n \sqrt{\det P}} e^{-\frac{1}{2}(y-\hat{y})^T P^{-1}(y-\hat{y})} \qquad \text{(I-11)}$$

with the mean $\hat{y}$ and the covariance matrix

$$P = \underset{y}{E}\{(y - \hat{y})(y - \hat{y})^T\} \quad , \qquad \text{(I-12)}$$

the entropy becomes

$$H = \frac{1}{2} \log [(2\pi e)^n \det P] \quad . \qquad \text{(I-13)}$$

It is seen that entropy is related rather directly to the covariance in this special case. This general connection holds also for non-Gaussian densities, but it is clear that the statistical parameter H can only provide a coarse summary of the probability density function $p(y)$.

D. SUMMARY OF RESULTS

The main results of the research performed in this study are

(1) A clear understanding was obtained about the required mathematical description of information.

   (*a*) The *quality* of the information-handling components (*i.e.*, sensors, communication links, and controller) of a system control, is determined by the probability density function of the state of the plant conditioned upon all available knowledge.

   (*b*) The *quantity* of information gained by improving the quality of the information-handling components can be measured by the average reduction in entropy of the conditional probability density function as a result of this improvement.

   (*c*) The *value* of the information gained by improving the quality of the information-handling components is measured by the average improvement in performance made possible by this additional information.

(d) Knowledge of the quantity of information is, in general, of little benefit to the design of control systems; knowledge of the value of information, on the other hand, is of fundamental importance.

(2) In order to determine performance, the combined optimization problem was formulated and solved for the general case of a dynamic process (I-1) subjected to random perturbations w_t and observed through an imperfect, *i.e.*, noisy and incomplete, measurement system (I-6). This formulation provides a theoretical answer to the central project objective, namely to relate the performance of control and guidance systems to the characteristics of the information links between the various constituents, under the assumption that the information fed to the controller is utilized optimally. In actual practice, the exact solution in the general case exceeds the computational capacities of present-day computers; however, many special cases have a computable solution and fairly evident approximations can be used. In addition, the solution is predicated on the assumption that the complete distributions $p(v_t)$ characterizing the measurement system and $p(w_t)$ characterizing the environment are known to the designer. The combined optimization problem is further discussed in Sec. III and Ref. 2.

(3) This general solution was applied to the important case of a linear system perturbed by Gaussian random effects and a concise mathematical expression was derived to relate system performance to the quality and structure of the measurement subsystem. This expression shows the trade-offs between system performance and the quality (and hence cost) of the various sensors and other information-handling components.

The main results of this special case are discussed in Sec. III and detailed derivations are given in Ref. 2.

(4) The general solution was particularized to the case of an "open-loop" system, where the decisions made on the basis of the noisy measurements do not affect the state of the dynamic process under observation. This has led to a general theory of optimum Information Systems, which is of great practical importance whenever the cost of collecting information is high compared to the cost of processing information. Section IV summarizes the results of this study, a detailed account of which is found in Appendix A.

(5) The relation between performance and the information-handling characteristics of the key constituents was derived for the general case of a fixed controller. This differs from the combined optimization problem in that the controller is not specifically designed to make optimum use of the measurements and, as a consequence, may be much simpler. This is the situation usually encountered in standard feedback control systems

where one wishes to know the magnitude of the degrading effects caused by noise upon a system structure selected to satisfy the design objectives in the absence of noise.

The problem was treated from the following two points of view:

(*a*) The conventional control engineering point of view based on well-known servo theory, notably the statistical methods of Newton, Gould, and Kaiser, sensitivity of performance to bias effects, and transient response of linear systems.

(*b*) The modern point of view based on the calculation of the probability density functions of the state in terms of the probability density functions of the random effects w and v. The exact and general solution was applied to the special case of a linear system perturbed by non-Gaussian effects. An example involving a linear plant and controller and a noisy digital feedback path was treated both by approximation of the Fokker-Planck equation and by empirical approaches verified by means of Monte Carlo simulation.

It was found that the entropy of the probability density of the state could not be updated from time $k + 1$ to time k without knowledge of the complete probability density at time k.

The results of these studies are given in Sec. VI-B and Appendix D for the conventional point of view and in Sec. VI-C and Appendices C and E for the modern point of view. The example problem involving a digital feedback path is treated in Sec. VI-D and Appendix F.

(6) The first problem attacked in the course of the study involved the optimum design of a quantizer located in the feedback path of an otherwise linear system. It was shown that for the important case of linear systems and quadratic performance measures, the optimization of the quantizer (the choice of the quantizer steps and switchpoints) could be performed separately from the optimization of the controller. This same result was shown to hold for the case of an imperfect measurement system where optimization of the quantizer, the controller, and the estimator can be carried out separately under justified simplifying assumptions. To optimize the quantizer, a convenient iterative

scheme replacing the customary and inefficient simultaneous search over many variables by a sequential search over one variable was developed. This useful design procedure is outlined in Sec. V and discussed in detail in Appendix B..

II INFORMATION MEASURES

In this section, the quantity of information (a measure of information based upon information-theoretic concepts) and the value of information (a measure of information based on control-theoretic concepts) are defined and related to the quality of information handling components such as sensors, communications links, and computers. The reader is also referred to an excellent paper by Marschak[5] for a discussion of these topics.

A. MATHEMATICAL MODELS OF CONTROL SYSTEMS

Before the relation between information and control systems can be developed, it is necessary to discuss suitable models of control systems.

1. Block Diagram

Figure II-1 is a block diagram showing the major elements which appear in a typical closed-loop control system. An open-loop control system has a similar block diagram except that the components connecting the state of the plant x^p to the controller are absent (*i.e.*, the components below the dotted line are missing).

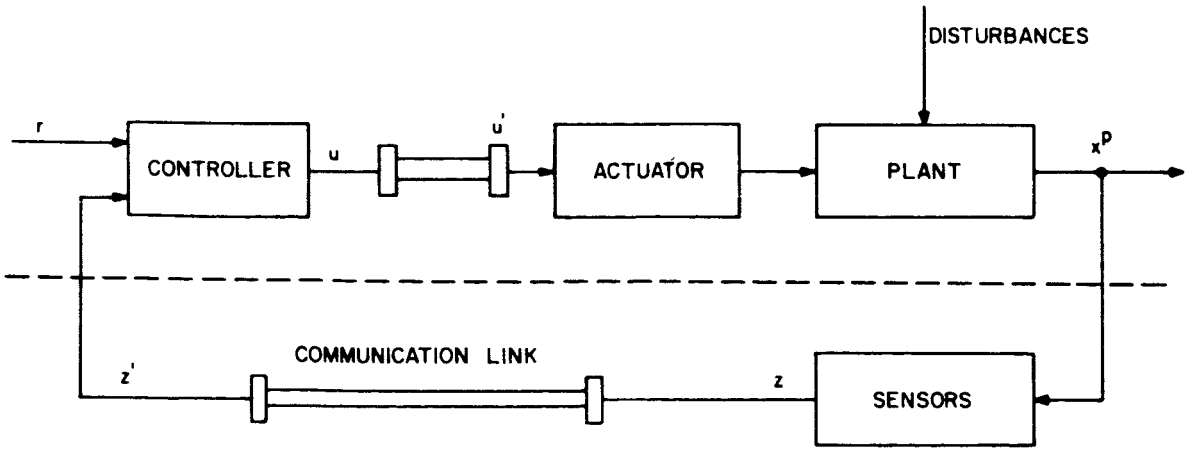

FIG. II-1 CLOSED-LOOP CONTROL SYSTEM

The classical method of analysis of such systems is the use of the Laplace transform theory with the various blocks represented by transfer functions.[6] In what follows, the modern state-space description of control systems is reviewed. It is desired to obtain a standard form by which control systems can be described.

2. Plant Description

The state x^p of the plant summarizes the effect of all past inputs to it as far as its future behavior is concerned. In other words, given the present state and all future inputs (including any random disturbance), the future behavior of the system can be determined. The state may be a vector of real numbers (in the case of ordinary dynamic systems), a vector of *zeroes* and *ones* (in the case of automata and "operational" systems) a function (in the case of distributed parameter systems) or a combination of these three.

Under the previously justified assumption of discrete time, the plant behavior can be described by

$$x^p_{k+1} = f^p(x^p_k, u_k, w^p_k, k) \qquad \text{(II-1)}$$

where w^p_k (over which the designer has no control) is the random input to the plant and where u_k (which must be selected by designer to achieve the desired performance) is the control input to the plant. Equation (II-1) is a relation telling how to determine the next state of the plant on the basis of the present state and present inputs.

Example:

Consider a rocket constrained to travel along a line and operating in a vacuum. If the mass of the rocket is fixed, then the present position and velocity summarize the effect of past inputs; hence, a suitable state is a vector with these two quantities as components. Application of Newton's laws generates the following state equation:

$$x^p_{k+1} = Ax^p_k + Bu_k + w^p_k \quad , \qquad \text{(II-2)}$$

where

$$A = \begin{bmatrix} 1 & \Delta t \\ 0 & 1 \end{bmatrix} \qquad B = \begin{bmatrix} (\Delta t)^2/2 \\ \Delta t \end{bmatrix}$$

$$x_k^p = \begin{bmatrix} \text{position at time } k \\ \text{velocity at time } k \end{bmatrix}$$

u_k = control thrust level for kth time interval

$$w_k^p = \begin{bmatrix} (\Delta t)^2/2 \\ \Delta t \end{bmatrix} \cdot \text{disturbance thrusts (assumed to be white)}$$

Δt = the time interval .

3. Sensor Description

The purpose of a sensor is to measure some property of the state of the plant. Because of imperfections, the output of the sensor will differ from the true value of the property being measured. Some major imperfections are:

(1) A bias error, an additive, time-independent—but often amplitude-dependent—departure; for example, gain changes of the sensor from nominal may be treated in such a manner.

(2) A noise-induced error, which is commonly assumed to be white (*i.e.*, uncorrelated from sample to sample). Noise not satisfying this assumption may be produced by use of proper filters operating upon white noise.

(3) An error caused by internal sensor dynamics.

(4) An error due to quantization and other nonlinearities.

Thus, a typical sensor is a dynamic system with an input that is a property χ of the state of the plant and with an output z, which is the measurement of χ. The sensor plant relation may be described quite generally by

$$\chi_k = g(x_k^p, k) \tag{II-3}$$

$$x_{k+1}^s = f^s(x_k^s, \chi_k, v_k^*, k) \tag{II-4}$$

$$z_k = h(x_k^s, \chi_k, v_k, k) \quad , \tag{II-5}$$

where x_k^s is the state of the sensor and v_k^* and v_k are measurement noises, assumed to be white.

Bias effects appear in this model as unknown values of initial conditions of the state x^s. Filtering of noise and internal dynamics are handled by the internal sensor state x^s. Quantization and other nonlinearities are reflected in the form of h.

Example:

Consider a position sensor with the following defects: a bias error consisting of an unknown zero offset and an unknown perturbation of the gain from a nominal value of 1, a delay in response consisting of a simple lag of time constant T, and a quadratic nonlinearity to the output. This sensor may then be described by (II-3) to (II-5), where

$$\chi_k = \text{position}$$

$$x_0^s = \begin{bmatrix} 0 \\ \text{offset} \\ \text{gain perturbation} \end{bmatrix}$$

$$f^s(x_k^s, v_k^*, \chi_k, k) = \begin{bmatrix} e^{-\frac{\Delta t}{T}} x_k^{s,1} + \left(1 - e^{-\frac{\Delta t}{T}}\right)(1 + x_k^{s,3})(\chi_k - x_k^{s,2}) \\ x_k^{s,2} \\ x_k^{s,3} \end{bmatrix}$$

$$h(x_k^s, v_k, \chi_k, k) = x_k^{s,1} + \alpha(x_k^{s,1})^2$$

T = time constant of the lag

α = known constant.

Note that components of vectors have been denoted by superscripts.

4. Communication Links

The purpose of communication links is to transfer information from one part of the system to another part, which may be separated by a considerable distance. Without going into details, it can be stated that communication links may also be represented by equations of the form (II-3) through (II-5).

5. Actuators

Actuators are devices for applying the control signals determined by the controller to the plant. Again, these components may be described by equations similar to those of a sensor.

6. Canonical Model

To show how to reduce the block diagram given in Fig. II-1 to a canonical form, it is sufficient to consider sensors only, since communication links and actuators may be handled in an exactly analogous fashion.

A new state

$$x = \begin{bmatrix} x^{p} \\ x^{s_1} \\ \vdots \\ x^{s_m} \end{bmatrix} \tag{II-6}$$

is defined, where the superscript s_i identifies the ith sensor. The combination of plant and sensors can now be described by two equations:

$$x_{k+1} = f(x_k, u_k, w_k, k) \tag{II-7}$$

$$z_k = h(x_k, v_k, k) \tag{II-8}$$

where

$$w_k^{p} = \begin{bmatrix} w_k^{p} \\ v_k^{*(1)} \\ \vdots \\ v_k^{*(m)} \end{bmatrix} \qquad v_k = \begin{bmatrix} v_k^{(1)} \\ \vdots \\ v_k^{(m)} \end{bmatrix}$$

$$z_k = \begin{bmatrix} z_k^{(1)} \\ \vdots \\ z_k^{(m)} \end{bmatrix}$$

$$f(x_k, u_k, w_k, k) = \begin{bmatrix} f^p(x_k^p, u_k, w_k^p, k) \\ f^{s1}[x_k^{s1}, g^{(1)}(x_k^p, k), v_k^{*(1)}, k] \\ \vdots \\ f^{sm}(\quad\quad\quad\quad\quad\cdot) \end{bmatrix}$$

$$h(x_k, v_k, k) = \begin{bmatrix} h^{s1}[x_k^{s1}, g^{(1)}(x_k^p, k)\ v_k^{(1)}, k] \\ \vdots \\ h^{sm}(\quad\quad\quad\quad\quad\cdot) \end{bmatrix} .$$

For simplicity it is assumed that $E[w_k, v_k^{(i)}] = 0$, although this is not necessary.

By following this procedure, *any* control system may be represented by the block diagram given in Fig. II-2. The plant is governed by Eq. (II-7), known as the *state equation*, and the measurement system is governed by Eq. (II-8), known as the *measurement equation*. Note that the state of the plant in Fig. II-2 includes not only the dynamics of the physical plant, but those of the sensors, communication links, and actuators as well. Furthermore, any nonwhite measurement noise appears as a white disturbance to the plant, which includes a suitable shaping filter.

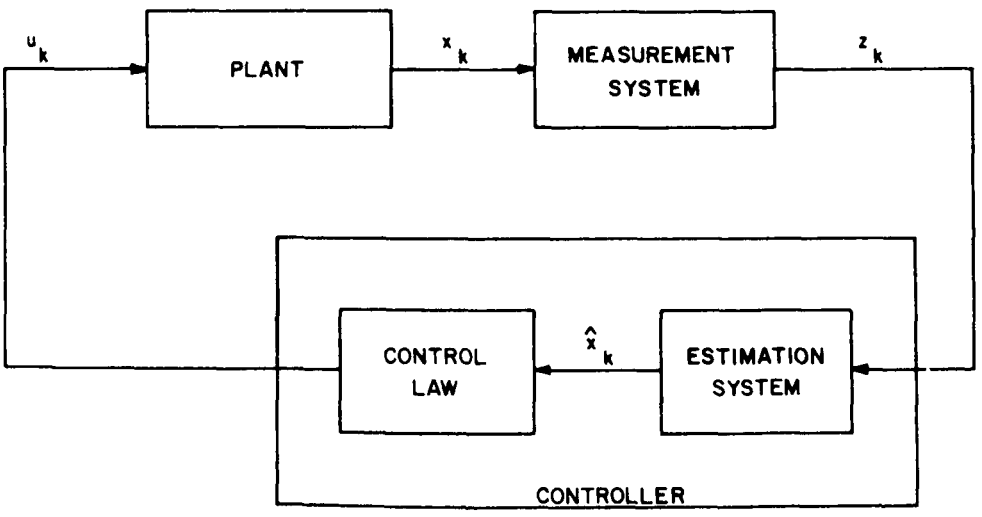

FIG. II-2 CANONICAL MODEL OF CONTROL SYSTEM OF FIG. II-1

To formulate a problem in the format given by Fig. II-2, it is necessary to carry out two steps. First, each component of the system must be modeled in the form given in Parts A-1 through A-5 above. Then these models must be reduced as described above. Modeling is often no easy task and must be based upon physical knowledge of the plant and the designer's experience as to what variables are important and what variables may be safely ignored. In many cases, testing is required to confirm the assumptions made.

To avoid confusion between the plant in Fig. II-1, which is a physical entity, and plant in Fig. II-2, which is a mathematical construction, the adjective *physical* will be used when referring to the actual plant. Similarly the term *sensors* will refer to Fig. II-1 and *measurement system* to Fig. II-2.

Example:

For the rocket ship described in Part A-2 and the position servo described in Part A-3, the overall state is

$$x_k = \begin{bmatrix} \text{position of rocket} \\ \text{velocity of rocket} \\ \text{internal dynamics of sensor} \\ \text{offset of sensor} \\ \text{gain perturbation of sensor} \end{bmatrix} .$$

7. Controller Description

Up to this point, the controller, which determines the inputs to be applied to the plant on the basis of available information about the plant, has been ignored. The controller can be broken into two parts: The estimation system is covered by

$$\hat{x}_{k+1} = \hat{f}(\hat{x}_k, z_k, k) \qquad \text{(II-9)}$$

and the control law by

$$u_k = \hat{h}(\hat{x}_k, z_k, k) \quad . \qquad \text{(II-10)}$$

The quantity $\hat{x}_k$, which in this report will be referred to as the *estimator of* x_k, is not in general what is thought of as the estimate of x_k. In particular it may have a dimension quite different from x_k. Any controller, optimal or suboptimal, may be represented in the form given

by (II-9) and (II-10); $\hat{x}_k$ is nothing more than the internal state of the controller in such a model. In Sec. III, in which the optimum $\hat{x}_k$ is found, the term estimator will become more clear.

Note that (II-9) and (II-10), which govern the controller, are similar to (II-7) and (II-8), which govern the remainder of the system. There is one notable difference: the presence of z_k in (II-10). This means that the present input to the controller can affect the present output of the controller. On the other hand, the present input of the plant cannot affect the present output of the measurement system; if it could, algebraic loops would be possible.

B. THE CONDITIONAL PROBABILITY DENSITY OF THE STATE AND THE QUALITY OF INFORMATION HANDLING

The state of the plant summarizes the past history of inputs insofar as they affect future behavior; hence, the control input to the plant can be based upon this quantity. Unfortunately, in the usual situation, the state of the plant is not available; therefore, a quantity that summarizes all knowledge about past inputs as they affect the future behavior must be sought.

In general, two sources of information about the state of the plant exist: the past control inputs and the past and present measurements. Thus, any prediction about the future behavior of the plant will be an expectation conditioned upon the values of these quantities. This expectation can be calculated from the state and measurement equations, the probability densities of the noises and disturbances, and the conditional probability density of the plant state given the past control inputs and measurements. The latter quantity is the only one which is not known *a priori**; hence, it appears intuitive that the control can be computed as a function of this quantity.

In Ref. 2 it is shown that the optimal control input can indeed be calculated as a function of the conditional probability density of state if the control problem is formulated in a suitable manner (see Sec. III). The $\hat{x}$ appearing in (II-8) and (II-9) is the conditional probability density: Estimation is calculation of the conditional probability density

* If the system equations or probability densities contain unknown parameters, the state can be augmented to include these parameters.

of the state, and control is calculation of the control input as a function of this conditional density.

Because it summarizes all past information and because it plays a role similar to the state of the plant in deterministic optimal control theory, the conditional probability density of the state will be referred to as the *information state* $\mathcal{P}_k$. There is one important difference between the state of the system and the information state. In the usual situation, the state of the plant is a finite dimensional vector, whereas the information state is a function and thus an infinite dimensional vector. In special cases the information state is finite dimensional; for example, if the conditional density is Gaussian, then it is completely specified by the conditional mean and the conditional variance of the state.

The purpose of the sensors, communication links, and of the estimator is to gather information about the state of the plant, transmit this information to the controller, and process it into a form suitable for making a decision by means of the control law. These components together may be termed the *information-handling system*. The quality of this system is determined by the spread of the conditional probability of the state, given the estimator $\hat{x}_k$. Note that in the general case, if the information transferral is to be perfect, then processing consists of calculating the conditional probability of state of the plant, given past measurements and inputs, since

$$p(x_k/\mathcal{P}_k) = p[x_k/p(x_k/Z_k, U_{k-1})] = p(x_k/Z_k, U_{k-1}) \qquad \text{(II-11)}$$

where U_k is used to represent the set $(u_0, u_1, \ldots, u_k)$ for an object u_i. Equation (II-11) shows that no information is lost in this processing.

A new element has been introduced into the model of a control system: To compute the information state, it is necessary to know the probability densities of the measurement noises and the disturbance inputs as well as the *a priori* probability density of the state of the plant. Knowledge of the physical processes involved in the generation of disturbances and noises may provide the form of these distributions and testing can be used to find their statistics. In many cases, however, translation of the designer's knowledge into the form needed for analysis is no easy task; this is particularly true for the *a priori* density of the state.

C. ENTROPY

Entropy, defined in the introduction, is the statistical quantity used in classical information theory as a measure of uncertainty.[7] The information provided by a given message about some quantity of interest is then defined as the amount of uncertainty removed by the message.

Since the information state $\mathcal{P}_k$ is a probability density, its entropy $\mathcal{H}$ may be calculated from

$$\mathcal{H}(Z_k, U_{k-1}, k) = -\int_{x_k} p(x_k/Z_k, U_{k-1}) \log p(x_k/Z_k, U_{k-1}) dx_k \quad . \quad \text{(II-12)}$$

In general, the value of U_{k-1} is not known *a priori* but is a random variable through its dependence upon the measurements. It is assumed that the controller is chosen in such a way that U_{k-1} is an optimum sequence in a sense yet to be specified. Then, it is possible from the *a priori* probability densities to calculate

$$\mathcal{H}(k) = E[\mathcal{H}(Z_k, U_{k-1}, k)] \quad . \quad \text{(II-13)}$$

$\mathcal{H}(k)$ is a measure of the expected uncertainty about the state of the plant at time k.

Let $\mathcal{H}_{OL}$ be the entropy about the state of the plant if the measurements are ignored, *i.e.*,

$$\mathcal{H}_{OL}(U_{k-1}, k) = -\int_{x_k} p(x_p/U_{k-1}) \log [p(x_k/U_{k-1})] dx_k \quad . \quad \text{(II-14)}$$

Similarly, let

$$\mathcal{H}_{OL}(k) = E[\mathcal{H}_{OL}(U_{k-1}, k)]$$

when optimum open-loop control is applied. Then

$$\mathcal{J} = \sum_{k=1}^{N} [\mathcal{H}_{OL}(k) - \mathcal{H}(k)] \quad \text{(II-15)}$$

is the amount of uncertainty about the state that is removed by the presence of the measurement system during the time up to N. $\mathcal{J}$ is the amount of information the measurement system provides about the state of the plant.

In a similar manner, the amount of information lost by use of imperfect communication channels and suboptimum information processing may be calculated.

Let $\mathcal{I}$ be calculated for two different sets of sensors, one of which has internal dynamics whereas the other does not. It is possible that $\mathcal{I}$ for the former system is larger than $\mathcal{I}$ for the latter simply because of information about the internal state of the sensors. Since the state of the physical plant in Fig. II-1 rather than the mathematical plant in Fig. II-2 is of primary interest, it is reasonable to replace $p(x_k/Z_k, U_{k-1})$ in the above equations by $p(x_k^p/Z_k, U_{k-1})$. It should be emphasized, however, that it is the former quantity that is needed for selecting the optimum control.

D. VALUE OF INFORMATION

To this point, the purpose of the information-handling system was taken to be to gain information about the state of the plant, and a measure of the amount of information gained was defined. No mention of the value of this information or the need for feedback control to gain more information than is *a priori* available was made. To answer these questions, the measures of the performance of control systems must be considered.

In classical control theory, performance measures were either of transient performance, such as pole location; or steady-state performance, such as error. Appendix C contains a discussion of the use of classical control theory and of modern control theory to analyze the effects of sensor imperfections on such classical control measures.

In what follows, variational performance measures will be considered. It is assumed that the cost of operating the system is given by

$$J' = \sum_{i=1}^{N} l[x(i), u(i), i] \qquad \text{(II-16)}$$

where one would like to minimize J'; however, because of the random effects present, one must settle for minimizing J, the expected value of J'.*

*J is often called a performance index and will be referred to in the sequel as the *performance*. It is unfortunate that cost rather than its negative profit was chosen, since to optimize performance it is necessary to minimize J.

While in some cases it is possible to specify the form of l from physical knowledge, in most cases selection of l is highly subjective.

Suppose that in order for the system to perform adequately, a level of performance J_{des} must be attained. Let J_{min} be the best performance obtainable with complete information about the plant and let J_{OL} be the best performance obtainable using only *a priori* information about the plant (*i.e.*, open-loop control). Then one of three possible cases may occur:

$$\begin{aligned} &(1)\quad J_{des} < J_{min} \leq J_{OL} \\ &(2)\quad J_{min} \leq J_{des} \leq J_{OL} \\ &(3)\quad J_{min} \leq J_{OL} < J_{des} \end{aligned} \qquad \text{(II-17)}$$

In Case (1), the system cannot perform adequately because the desired performance is better than the plant is capable of giving. In Case (3), the *a priori* information is sufficient and there is no need for sensing devices to gather information about the plant. In Case (2), the customarily encountered situation, it is necessary to make measurements on the plant in order to attain the desired level of performance.

The quantity $J_{OL} - J_{min}$ is an upper bound on the performance improvement resulting from perfect information. If J_M is the optimal performance with a given set of sensors, then $J_{OL} - J_M$ is the *value* associated with the use of that set. Also, if M_1 and M_2 are two alternative measurement systems for which

$$J_{M_1} = J_{M_2} = J_{des} \quad , \qquad \text{(II-18)}$$

then, by comparison of the dollar cost of these systems, the least expensive set can be chosen.

To calculate such quantities as J_{min}, J_{OL} and J_M, it is necessary to find the controller that minimizes the expected value of J for various measurement systems. The solution of this problem is given by combined optimization, considered in detail in Ref. 2 and in Sec. III. In effect, the solution of the information handling problem has been reduced to the solution of the optimum information utilization problem—that is, calculation of the best performance possible with a given system.

In a similar manner, the cost of using suboptimal processing and imperfect communication links is measured by the resulting degradation in performance. Thus, the value of more sophisticated processing and better communication channels can be computed to determine whether such an effort is worth the expense involved.

E. COST OF INFORMATION

There is no direct relationship between the amount of information and the value of that information, except in special cases, such as Kelly's gambler.[5,8] This is because in the information-theoretic sense, information is gained when uncertainty is reduced; however, this information may be completely irrelevant to the task to be performed. On the other hand, there is sometimes a closer relation between the amount of information and the dollar cost of collecting this information.

The relation between amount and dollar cost of information is most evident in the case of communication channels. By Shannon's theorem, the maximum rate at which information can be sent over a channel is equal to the capacity of the channel. The cost of building a channel in turn is a monotonic and often a linear function of the capacity.

For example, the capacity C of a binary communication channel, is (see Sec. VI and Appendix F)

$$C = 1 + p \log_2 p + (1 - p) \log_2 (1 - p) \qquad \text{(II-19)}$$

where p is the probability of error. Since it costs money to reduce the probability of error, the cost of the channel increases with the rate of information transfer.

While there is no close relation between the dollar cost of a measurement system as a whole and the amount of information it gathers, such a relationship does hold in the case of individual sensors. For example, to gain more information in a sensor with quantized output, it is necessary to increase the number of quantum levels, which in turn increase the cost, though not necessarily in proportion to entropy.

It is much harder to find a relationship between the cost of processing information and the amount of information processed. Information processing can be viewed as removing extraneous information from a signal to leave only the desired information. It is not unreasonable to assume that

the difficulty (and hence the cost of processing information) grows with the amount of extraneous information.

For example, in the linear case discussed in Sec. III-C, the optimum estimator is the conditional mean of the state. The measurements contain unwanted information about the measurement noise that is removed or filtered out by means of the estimation system, which in this case is the Kalman filter. In Ref. 2, examples are presented that indicate that the more extraneous information (*i.e.*, measurement noise) present, the more difficult it is to remove the unwanted information in the sense that more wanted information is lost by suboptimal processing.

While these intuitive relations between amount of information and cost of information exist and are important from a conceptual point of view, it does not appear that calculation of the expected entropies as described in Sec. II-C will be of much practical value in the design of control systems. The cost of information handling components is known directly and it is the value of using these components that is of interest This value is found by solving the combined optimization problem presented in Sec. III and in Ref. 2.

F. ILLUSTRATION

When measurements are made upon a plant in order to obtain adequate performance, several important questions arise, namely:

(1) What measurements should be made?

(2) How good do these measurements have to be?

(3) How can *a priori* and gathered information be used to simplify the information processing?

(4) How does one derive optimum control decisions for a given set of measurements?

The answer to the fourth question is provided by the combined optimization problem discussed in Sec. III and in Ref. 2. In this section, a simple illustration of how the solution of the combined optimization problem in conjunction with the concepts developed so far may be used to answer the other three questions is given.

In the important linear case, the solution of the combined optimization problem is well known (see Sec. III-D). In this case, the state and

measurement equations are both linear, the random disturbances and noises are Gaussian, and the performance criterion is quadratic. For simplification, it will be assumed that the system is time invariant and only steady-state performance will be considered. The optimum estimator for this case (*i.e.*, $\hat{x}$ in Fig. II-2) is the conditional mean of the state of the plant given all available information. The cost per time increment, defined by

$$\Delta\beta = \lim_{N\to\infty} \frac{\sum_{i=1}^{N} l(x_i, u_i, i)}{N} \quad , \tag{II-20}$$

is

$$\Delta\beta = \text{tr}\ [\hat{P}Q + P^* V] \quad , \tag{II-21}$$

where tr $[A]$ is the sum of the diagonal elements of the matrix A, $\hat{Q}$ is the covariance matrix of the disturbance noise, V is the covariance matrix of the error in the estimate, P and P^* are matrices found in the solution of the control problem and defined in Sec. III-D. This formula also holds for a very general class of suboptimal estimation procedures as long as optimal control is used.

Following the notation given in Sec. II-D, we note that

$$\Delta\beta_{min} = \text{tr}\ [\hat{P}Q]$$

$$\Delta\beta_{OL} = \text{tr}\ [\hat{P}Q + P^* V_{OL}] \quad , \tag{II-22}$$

where V_{OL} is the covariance in estimate of the state given no measurements. Hence, the maximum gain in performance per unit time that can be obtained with measurements is tr $[P^* V_{OL}]$.

The covariance V completely characterizes the quality of a given measurement and estimation system, since the difference in performance between two alternative systems can be determined purely on the basis of the covariances V_{M_1} and V_{M_2} corresponding to each. Furthermore, if $V_{M_1} > V_{M_2}$ (*i.e.*, if $V_{M_1} - V_{M_2}$ is positive definite) then, since $P^* \geq 0$, $\Delta\beta_{M_1} > \Delta\beta_{M_2}$. Reduction of the error in estimate of any one of the state variables by a given amount will result in an equal gain of information in terms of entropy; however, the changes in performance may differ greatly depending upon P^*. Thus, quantity of information and value of information are not necessarily closely related. Furthermore, improvement in the quality of

the information-handling components provides more information but does not necessarily improve performance correspondingly.

Those state variables of the plant that correspond to large elements of P^* should be known well for high performance. This, however, does not imply that all these state variables must be measured; it may well be that measurement of one state variable provides the necessary information about another system. In fact, if the plant is observable[9] with the given measurement system, then the measurements will contain some information about the whole state. If this information is not sufficient for adequate control, two alternatives are available for gaining additional information: adding additional sensors or improving the existing sensors. Hence, the answers to Questions (1) and (2) raised at the beginning of this part are determined by calculating P^* and V. In a similar manner P determines which state variables are most sensitive to disturbance inputs.

For example, in a position control system, the position sensor provides information about rate, which in many cases will be adequate to achieve the desired performance. If the available information is not sufficient, the position sensor may be improved enabling better calculation of rate; or rate may be measured directly.

In general, one of three actions can be taken for a state variable:

(1) Measure the state variable.

(2) Compute the state variable from measurements of other state variables (if the resulting V permits).

(3) Ignore the state variable (if the P^* permits).

Suboptimal processing, in particular suboptimal estimation, will now be considered. If, for a given measurement system, V_M is the covariance of the error in estimate with optimal estimation and V_M' the covariance with suboptimal processing, then $V_M' > V_M$, the degradation in performance is given by

$$\text{tr}\ [P^*(V_M' - V_M)] \quad . \tag{II-23}$$

If V_M is much smaller than is necessary for adequate performance, either because of *a priori* knowledge or because of very good measurements, then a fair degree of suboptimal processing may be tolerated and adequate performance still attained.

In general, there may be many alternative ways to obtain the same performance for a given plant. By comparing the dollar cost of building alternative systems, one can choose the most economical system for obtaining the desired performance. The combined optimization problem is the problem of calculating performance for such systems. In Ref. 2, several numerical examples are presented to illustrate the various alternative ways for obtaining equivalent performance in the linear case. Some of the trade-offs considered are

(1) Number of sensors versus accuracy of sensors.

(2) Quality of the measurement system versus "isolation" from disturbances.

(3) Quality of the measurement system versus quality of the estimation system.

(4) Quality of *a priori* information versus quality of the estimation system.

III COMBINED OPTIMIZATION PROBLEM

In the previous section, we saw that it is necessary to solve the combined optimization problem in order to calculate the value of information in a control system. In this section, the theory of the combined optimization problem is presented briefly; the theory is presented in detail in Ref. 12, which contains a list of references to other work on the problem. One excellent reference on the subject is by Wonham.[10]

A. PROBLEM STATEMENT

For convenience, the combined optimization problem, which was presented in bits and pieces in the previous chapter, is restated in its entirety here. The problem is illustrated in Fig. II-2.

Given (1) A plant, described by

$$x_{k+1} = f_k(x_k, u_k, w_k) \qquad \text{(III-1)}$$

where

x_k is the state vector

u_k is the control or input vector

w_k is the disturbance vector, assumed to be white.

(2) A measurement system, described by

$$z_k = h_k(w_k, v_k) \qquad \text{(III-2)}$$

where

z_k is the measurement vector

v_k is the measurement noise vector, assumed to be white.

(3) The probability distributions

$$\text{(a)} \quad p(x_0) \tag{III-3a}$$

$$\text{(b)} \quad p(w_i) \qquad i = 0,\ldots,N \tag{III-3b}$$

$$\text{(c)} \quad p(v_i) \qquad i = 0,\ldots N \tag{III-3c}$$

(4) The performance index

$$J = E\left\{\sum_{i=0}^{N} l(x_i, u_i, i)\right\} \tag{III-4}$$

(5) The admissibility constraint

$$u_i \in \Omega_i \tag{III-5}$$

Find the admissible controller that minimizes J, where

(1) A controller is defined as any algorithm that at time k generates u_k as a function of the present and all past measurements $(z_k,\ldots,z_0)$.

(2) An admissible controller is defined as any controller which, when used in the closed-loop system shown in Fig. II-2, yields admissible u_i.

B. SIMPLIFIED DERIVATION OF SOLUTION

In this part, we present a simplified derivation of the solution of the combined optimization problem.

1. Stochastic Control Problem

Before the general problem is treated, we consider the solution of the special case in which the measurement system is perfect (*i.e.*, $z_k = x_k$). For this case, no estimation is necessary; it is referred to as the *stochastic control problem**. Bellman has derived a recursive equation for solution of this problem[3]: If $I(x_k,k)$ is defined by

$$I(x_k,k) = \min_{u_k(x_k),\ldots,u_N(x_N)} E\left[\sum_{i=k}^{N} l(x_i,u_i,i)\,\middle|\,x_k\right] \tag{III-6}$$

*Wonham[10] and others have referred to the combined optimization problem defined in A as the stochastic control problem; we prefer Bellman's more restrictive definition.[3]

then $u_k(x_k)$ is found by solution of

$$I(x_k,k) = \min_{u_k}\left(l(x_k,u_k,k) + \underset{w_k}{E}\{I[f_k(x_k,u_k,w_k),k+1]\}\right) \qquad k < N$$

$$I(x_N,N) = \min_{u_N} l(x_N,u_N,N) \qquad \text{(III-7)}$$

where the w_k under the E indicates that the expectation is to be taken with respect to x_k.

2. CONTROL EQUATION

From the discussion presented in Sec. II-B it seems intuitively obvious that the control u_k at time k can be chosen as a function of ρ_k, the conditional probability density of the state given past inputs u_{k-1} and past and present measurements z_k. In.Ref. 2, this dependence is not assumed, but proved rigorously. In this section we assume that u_k is a function of ρ_k and then solve the combined optimization problem by converting it into a stochastic control problem.

In the next section we show that ρ_k satisfies the recursion relation

$$\rho_{k+1} = F_k(\rho_k, u_k, z_{k+1}) \quad . \qquad \text{(III-8)}$$

Equation (III-8) has the form of a state equation for a plant with state ρ_k, input u_k and disturbance z_{k+1}.

Consider the performance index in Eq. (III-4). By use of a well-known identity on conditional expectations[11] each term of the summation may be written

$$E[l(x_i,u_i,i)] = E\underset{x_i}{E}\left(\{[l(x_i,u_i,i)/\rho_i]\}\right)$$

$$= E[L(\rho_i,u_i,i)] \quad , \qquad \text{(III-9)}$$

where

$$L(\rho_i,u_i,i) \triangleq \underset{x_i}{E}\,[l(x_i,u_i,i)/\rho_i]$$

$$= \underset{x_i}{E}\,[l(x_i,u_i,i)/Z_iU_{i-1}] \qquad \text{(III-10)}$$

Hence we may rewrite (III-4 as)

$$J = E\left[\sum_{i=0}^{N} L(\rho_i,u_i,i)\right] \qquad \text{(III-11)}$$

Comparing Eqs. (III-8) and (III-11) with Eqs. (III-1) and (III-4) we see that the former two equations define a stochastic control problem in which $\mathcal{P}_k$ is the state, z_{k+1} is the disturbance and u_k is the control. This problem differs from the ordinary stochastic control problem in that the state is not finite dimensional; however, in the derivation of Eq. (III-7) no property of the dimension of x_i is used. Hence we may write down at once

$$I^*(\mathcal{P}_k,k) = \min_{u_k}\left(L(\mathcal{P}_k,u_k,k) + \underset{z_{k+1}}{E}\left\{I^*\left[F_k(\mathcal{P}_k,u_k,z_{k+1}),\ k+1\right]\right\}\right) \qquad k < N$$

$$I^*(\mathcal{P}_N,N) = \min_{u_N} L(\mathcal{P}_N,u_N,N) \qquad \text{(III-12)}$$

where

$$I^*(\mathcal{P}_k,k) \triangleq \min_{u_k} E\left[\sum_{i=k}^{N} L(\mathcal{P}_i,u_i,i)/\mathcal{P}_k\right],$$

which is identical with Eq. (II-32) of Ref. 2. Because solution of Eq. (III-12) enables one to determine the optimum control law (*i.e.*, to determine u_k as a function of $\mathcal{P}_k$), we will refer to it as the *control equation*.

3. Estimation Equation

In this section we present the form of the F given in Eq. (III-8); in keeping with the notation of Sec. III-A this equation will be referred to as the *estimation equation*. It may be used to update the probability density in real time; alternatively, we may use Eq. (III-8) in conjunction with Eq. (III-12) to determine the controls u_k as a function of z_k rather than $\mathcal{P}_k$. These two methods of specifying control are treated in detail in Ref. 2.

An equation for updating the conditional probability of the state of a plant with no control inputs is derived by application of Bayes rule in Appendix A. This equation may be modified to handle control inputs by simply replacing $p(x_{k+1}/x_k)$ by $p(x_{k+1}/x_k,u_k)$. The result is

$$p(x_{k+1}/Z_{k+1},U_k) = \frac{p(z_{k+1}/x_{k+1})\int_{x_k} p(x_{k+1}/x_k,u_k)p(x_k/Z_k,U_{k-1})dx_k}{\int_{x_{k+1}} p(z_{k+1}/x_{k+1})\int_{x_k} p(x_{k+1}/x_k,u_k)p(x_k/Z_k,U_{k-1})dx_k dx_{k+1}}$$

$$p(x_0/Z_0,U_{-1}) = \frac{p(z_0/x_0)p(x_0)}{\int_{x_0} p(z_0/x_0)p(x_0)dx_0} \qquad \text{(III-13)}$$

Since $\mathcal{P}_k \triangleq p(x_k/Z_k,U_{k-1})$, Eq. (III-13) indeed specifies the form of F in Eq. (III-8).

4. Comments

Because the combined optimization problem is equivalent to an infinite dimensional control problem, we can expect considerable difficulties in calculating exact solutions in general. Suitable approximations are proposed in Ref. 2; investigation of approximations is a major area of unfinished research on even the classical optimal control problem, not to mention the combined optimization problem.

In two special cases, however, the combined optimization problem is not infinite dimensional: If Eqs. (III-1) and (III-2) are linear; if Eqs. (III-3a), (III-3b), and (III-3c) are Gaussian; and if Eq. (III-4) is quadratic; then Eqs. (III-12) and (III-13) reduce to a well-known matrix equation. If x_k can take only a finite number of values, then p_k is just the vector of probabilities that the plant is in one of its possible states and Eqs. (III-12) and (III-13) become finite sets. The solution to the linear case is summarized in Part C; Ref. 2 presents the detailed derivation of the results as well as a list of the many references in which the solution is given. Part D discusses an example of a finite state system. Solution of Eq. (III-13) requires knowledge of $p(x_{k+1}/x_k,u_k)$ and $p(z_k/x_k)$. The first of these probability densities can be obtained from the state equation and $p(w_k)$; $p(x_{k+1}/x_k,u_k)$ is an alternative way of describing a randomly disturbed plant. Similarly, $p(z_k/x_k)$ can be obtained from knowledge of the measurement equation and $p(u_k)$ and is an alternative way of describing the measurement system.

C. LINEAR CASE

In this part, a very important special case of the combined optimization problem is considered.

1. Statement

The linear case of the combined optimization problem refers to the situation in which the following assumptions hold:

(1) The plant and measurement systems are linear, *i.e.*,

$$\text{(a)}\quad x_{k+1} = A_k x_k + B_k u_k + w_k \tag{III-14}$$

$$\text{(b)}\quad z_k = C_k x_k + v_k \quad .$$

(2) The performance index is quadratic, *i.e.*,

$$l(x_i, u_i, i) = x_i^T Q_i x_i + u_i^T R_i u_i \tag{III-15}$$

(3) The probability distributions are Gaussian, *i.e.*,

$$\text{(a)}\quad p(x_0) = c_1 \exp\left[(x_0 - \bar{x}_0)^T (\hat{Q}_{-1})^{-1} (x_0 - \bar{x}_0)\right] \tag{III-16}$$

$$\text{(b)}\quad p(w_k) = c_2 \exp\left(w_k^T \hat{Q}_k^{-1} w_k\right) \tag{III-17}$$

$$\text{(c)}\quad p(v_k) = c_3 \exp\left(v_k^T \hat{R}^{-1} v_k\right) \tag{III-18}$$

where c_1, c_2, c_3 are constants of no consequence here and where:

$\hat{Q}_{-1}$ = *a priori* covariance of x_0

$\hat{Q}_k$ = covariance of the disturbance at time k

$\hat{R}_k$ = covariance of the measurement noise at time k

$\bar{x}_0$ = *a priori* mean of x_0 .

2. Solution

In the linear case the controller takes the form given in Fig. III-1. Note that all the matrices are given except G_k and K_k, which are found by solution of the control problem and estimation problem respectively.

For any controller with the form given in Fig. III-1, but not necessarily with optimum G_k and K_k;

$$J = \bar{x}_0^T P_0 \bar{x}_0 + \text{tr}\,(P_0 \hat{Q}_{-1}) + \sum_{k=0}^{N-1} \Delta\beta_k \tag{III-19}$$

where

$$\Delta\beta_k = \text{tr}\left[P_{k+1} \hat{Q}_k + P^*_{k+1} V_k + 2P'_k K_k (\hat{R}_k K_k^T - C_k V_k)\right] \tag{III-20}$$

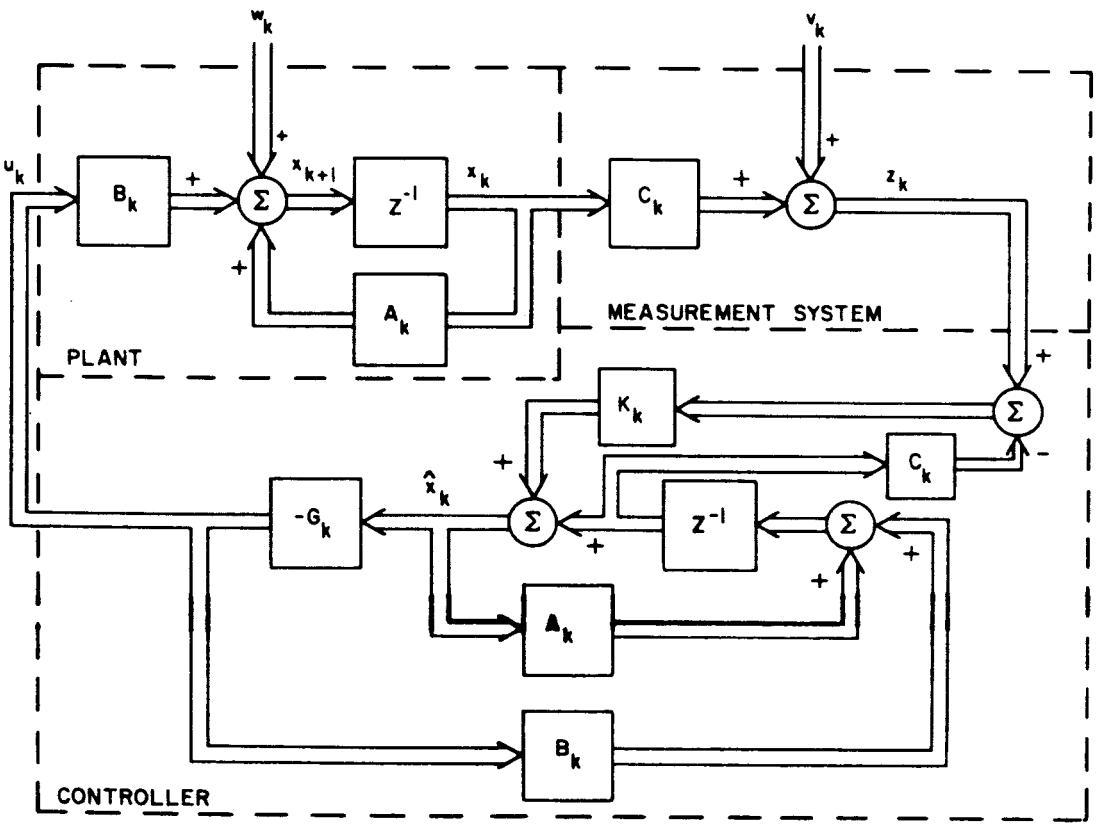

FIG. III-1 LINEAR COMBINED CONTROL AND ESTIMATION SOLUTION

$$P_k = Q_k + G_k^T R_k G_k + (A_k - B_k G_k)^T P_{k+1}(A_k - B_k G_k) \qquad k = 0,1,\ldots,N-1 \tag{III-21}$$

$$P_N = Q_N$$

$$P_{k+1}^* = Q_k + A_k^T P_{k+1} A_k - P_k \tag{III-22}$$

$$V_{k+1} = \tilde{Q}_k + (I - K_{k+1}C_{k+1})A_k V_k A_k^T (I - K_{k+1}C_{k+1})^T \qquad k = 0,1,\ldots,N-1 \tag{III-23}$$

$$V_{-1} = \tilde{Q}_{-1}$$

$$\tilde{Q}_k = K_{k+1}\hat{R}_{k+1}K_{k+1}^T + (I - K_{k+1}C_{k+1})\hat{Q}_k(I - K_{k+1}C_{k+1})^T \tag{III-24}$$

$$(I - K_k C_k)P_k' = (A_k - B_k G_k)^T[(I - K_{k+1}C_{k+1})P_k'(I - K_{k+1}C_{k+1}) - P_{k+1}^* B_k G_k] + G_k^T R_k G_k \qquad k = 0,1,\ldots,N-1$$

$$P_N' = 0 \tag{III-25}$$

Note that $\Delta\beta_i$ is the cost of operating during the interval $[k, k+1]$.

The optimum G_k and K_k are given by

$$G_k = (B_k^T P_{k+1} B_k + R_k)^{-1} B_k^T P_{k+1} A_k \tag{III-26}$$

and

$$K_k^T = \hat{R}_k^{-1} C_k V_k \tag{III-27}$$

For optimum control, Eqs. (III-21) and (III-22) reduce to

$$P_k = Q_k + A_k^T P_{k+1} A_k - A_k^T P_{k+1} B_k (B_k^T P_{k+1} B_k + R_k)^{-1} B_k^T P_{k+1} A_k \qquad k = 0,1,\ldots N-1$$

$$P_N = Q_N \tag{III-28}$$

and

$$V_k = \hat{P}_k - \hat{P}_k C_{k+1}^T (C_{k+1}\hat{P}_k C_{k+1}^T + \hat{R}_{k+1})^{-1} C_{k+1}\hat{P}_k \tag{III-29}$$

where

$$\hat{P}_{k+1} = \hat{Q}_k + A_{k+1}\hat{P}_k A_{k+1}^T - A_{k+1}\hat{P}_k C_{k+1}^T (C_{k+1}\hat{P}_k C_{k+1}^T + \hat{R}_{k+1})^{-1} C_{k+1}\hat{P}_k A_{k+1}^T \qquad k = 0,1,\ldots,N-1$$

$$\hat{P}_{-1} = \hat{Q}_{-1} \tag{III-30}$$

In essence, Eqs. (III-28) and (III-30) are the control equation and the estimator equation respectively; therefore, for the linear case, the control equation and estimation equation are finite dimensional equations. Furthermore, the two equations have the same form mathematically; this result is Kalman's duality principle. When the optimum K_k is used the estimator is the Kalman-Bucy filter[9], the time varying, multivariable extension of the Wiener filter.

If optimum control is used [*i.e.*, Eq. (III-26) holds], then P_k' is zero. On the other hand, if optimum estimation is used [*i.e.*, Eq. (III-27) holds], then the last two terms of Eq. (III-20) cancel.

In either case,

$$\Delta\beta_k = \mathrm{tr}\,[P_{k+1}Q_k + P^*_{k+1}V_k] \quad . \qquad \text{(III-31)}$$

All of the results presented here are derived in Ref. 2; however, since they are scattered in that report, this section presents a compact summary of the solution to the linear case. Most of the results for the optimal processing are well known (see References in Ref. 2); however, the results for suboptimal processing appear to be original. A formula translator ASP (Automatic Synthesis Program) has been developed for NASA by Kalman and Englar[12] for programming equations such as those presented in this section.

D. DUAL CONTROL

In Sec. II, it was shown that the purpose of feedback control was to gain information about the state of the plant in the absence of sufficient *a priori* information. In addition to the use of sensors to gather information, and filters to process information, there is a third method of gaining usable information about the state of the plant: control action. Intuitively, the concept of using control action to gain information is nothing more than using test signals as inputs in order to gain information; hence, the control input can be used to gain information as well as control the plant. Since the input that provides the most expected information is not necessarily the one which is most likely to bring the plant to a desired state, it is necessary to consider the compromise between the control and informational uses of the control input. In recognition of the dual purpose of control inputs, Feldbaum[13] refers to the problem as Dual Control.

The most obvious application of dual control theory is to adaptive control systems—indeed this was the original motivation behind Feldbaum's work; however, control action can be used to help estimate the state of a plant in a nonadaptive situation as the example of dual control in Ref. 2 shows.

In the linear case, the uncertainty about the state is completely determined by the covariance matrix. Since this matrix is unaffected by the controls applied, no dual control need be considered. Thus, to present a simple example of dual control, it may be necessary to go to the other special cases—discrete state systems—in which the optimal estimator is finite dimensional. For the present purposes, it is sufficient to paraphrase the example presented in Ref. 2. To make the example more stimulating, it has been given an admittedly contrived physical interpretation.

The plant under consideration is a spacecraft with a Jovian fly-by mission. The spacecraft attitude is maintained by the sun and a reference star. Attitude lock of the spacecraft is accomplished by means of an acquisition procedure; however, if it is already in lock when this procedure is applied, the system will be thrown out of lock. The system can perform a test to see if it is locked on; the test never indicates lock when it does not exist but in a manner independent may fail from time to time to indicate lock when it actually exists. It is assumed that, once locked, the system remains in lock unless the acquisition procedure is applied.

For this system, a valid information state is the probability that the system is in lock. Suppose that probability is less than 0.5; then the optimum input from a control viewpoint is to instigate reacquisition. On the other hand, the optimum input from the information point of view is to leave the system alone since, if the assumption that the system is out of lock is wrong, the test will discover this fact sooner or later.

Now suppose that in transit there is very little cost if the spacecraft is not attitude-stabilized, but that it is very important that it be stabilized during the last time increment (*i.e.*, the fly-by). If at the time of the next to last control input (*i.e.*, $N-2$) the information about the state is poor then it is profitable to take that action which provides the most information so that the last control input u_{N-1} can be made with most probability

of resulting in lock during fly-by $[(N - 1)\,\Delta t$ to $N\Delta t]$. Thus, in this case, the optimal input for the next to last time increment is that which gains most information.

Now consider the same system and situation, except that the test is performed to determine lack of lock. In this situation, the test will never indicate lock when it does not exist; however, it may fail to indicate lack of lock. For probability of lock less than 0.5, the optimum action from a control point of view is to instigate reacquisition; furthermore, in this case, this action also provides most information. It is clear that this control is optimum in an overall sense since, if the decision made is wrong, the test will sooner or later determine this fact.

Thus, two situations have been considered that are essentially the same except for what amounts to a change in polarity of the measurement system. In one case, the best controls from both the control and information viewpoints are the same; in the other case, they differ. As would be expected, the average performance in the case where a dual control compromise must be made is less than the case where no such compromise is necessary. Whenever possible, it pays to arrange the measurement system so that the control most likely to take the system to the desired state also provides most information.

E. EXTENSIONS OF THE THEORY

In this part, it is shown how several problems that at first do not appear to be combined optimization problems can be formulated in a manner such that they become combined optimization problems.

1. Augmentation of the State

In Sec. III-A, it was shown that, by augmenting the state to include the dynamics of the sensor and other components, the general control system could be reduced to the canonical form given in Fig. II-2. This technique can also be applied to reduce a number of problems to combined optimization theory.

a. ADAPTIVE CONTROL

If there are unknown parameters of either the physical components or the random disturbances and noise, these may be handled by converting the parameters to state variables. From this point of view, an adaptive

control problem is a nonlinear control problem. If the unknown parameters are randomly time varying, then it will be necessary to add state variables to account for this fact. An alternative to specifying initial distributions of the unknown parameters is to assume that nature is trying to pick the parameters in the most deleterious manner, a problem which can be treated by the theory of differential games.

b. CONTROLLABLE PARAMETERS

In many situations, it may be possible to change the mode of operation of components; for example, a radar may be operated either in a scan or a track mode. One method of making optimum decisions about the mode of operation is to augment the state to include the mode of operation and to augment the control input to allow changes in the mode of operation.

c. SUM-TYPE CONSTRAINTS

In the problem formulation, the only type constraints permitted were instantaneous constraints; however, many constraints of importance are of the form

$$\sum_{i=0}^{N-1} G(u_i, i) \leq C \quad . \tag{III-32}$$

For example, if the plant is a rocket, the summation could be the total fuel used and C the fuel available.

Define a new state variable by

$$\begin{aligned} x'_{k+1} &= x'_k + G(u_k, k) \\ x'_0 &= 0 \end{aligned} \tag{III-33}$$

then

$$\sum_{i=0}^{N-1} G(x_i, u_i, i) = x'_N \tag{III-34}$$

and (III-32) becomes

$$x'_N \leq C \quad . \tag{III-35}$$

This constraint is an instantaneous state variable constraint, which was not included in the original formulation. However, since x_k' can be measured directly, such a constraint may be handled in the same manner as instantaneous constraints upon u.

An alternative method of handling such constraints is the Lagrange multiplier, but in many situations this method is not valid.

2. Suboptimal Performance

In many cases, it is desirable to calculate the performance of a control system (Fig. II-2) with a fixed but not necessarily optimal controller. Section IV and Appendices D through F present a detailed discussion of fixed controller systems; in this paragraph, a method of calculating the performance by use of combined optimization theory is presented.

Assume that the controller is governed by Eqs. (II-9) and (II-10). Consider the whole system of Fig. II-2 as a plant to be controlled with state

$$x_k' = \begin{bmatrix} x_k \\ \hat{x}_k \end{bmatrix}. \tag{III-36}$$

From (II-7), (II-8), (II-9), and (II-10), this plant is governed by the state equation

$$\begin{aligned} \hat{x}_{k+1}' &= f'(x_{k+1}', w_k') \\ &= \begin{bmatrix} f\{x_k, \hat{h}[\hat{x}_k, h(x_k, v_k, k), k], w_k, k\} \\ \hat{f}[\hat{x}_k, h(x_k, v_k, k), k] \end{bmatrix}, \end{aligned} \tag{III-37}$$

where

$$w_k' \triangleq \begin{bmatrix} w_k \\ v_k \end{bmatrix}. \tag{III-38}$$

Let the measurement system for this plant be

$$z_k' \triangleq x_k' \tag{III-39}$$

and the performance measure be

$$J' = E \sum_{i=0}^{N} \left(l\{\hat{x}_i, \hat{h}[\hat{x}_i, h(x_i, v_i, i), i], i\} + u_i'^2 \right)$$

$$= J + E\left\{ \sum_{i=0}^{N} u_i'^2 \right\} \qquad \text{(III-40)}$$

Equations (III-37), (III-39), and (III-40) together with the statistics of v_k and w_k, define a combined optimization problem. This problem is of a very special kind, however: The control input u_i' has no effect on the next state. Therefore, from (III-40), it is obvious that the optimal u_k is given by

$$u_k' \equiv 0 \qquad \text{(III-41)}$$

and that the optimal performance

$$J'_{MIN} = J \quad . \qquad \text{(III-42)}$$

Hence, by solution of the combined optimization problem described by (III-37), (III-39), and (III-40), the performance of the original fixed system is found, but since by (III-41) $u_k \equiv 0$, no minimization is then necessary to the solution of this problem.

In Ref. 2, this method plus considerable algebraic manipulation is used to derive the results presented in Part C for nonoptimal controllers in the linear case.

IV OPTIMUM INFORMATION SYSTEMS

Optimum information systems are distinguished from combined optimization problems, as discussed in Sec. III, by the fact that the decisions u_k made on the basis of the observations z_k do not affect the signal generating process. Hence, they constitute a special case of combined optimization theory, and in addition to their very real practical importance, provide another and somewhat simpler vehicle for discussing and substantiating the concepts of *quantity* of information and *value* of information.

The work on optimum information systems was greatly stimulated by J. Marschak's paper[5]. In Appendix A, Marschak's qualitative discussion is complemented by quantitative expressions and his problem formulation is considerably expanded by modeling the signal generating process and by considering a much more general cost function.

A. PROBLEM FORMULATION

With reference to Fig. IV-1, the problem is defined as follows:

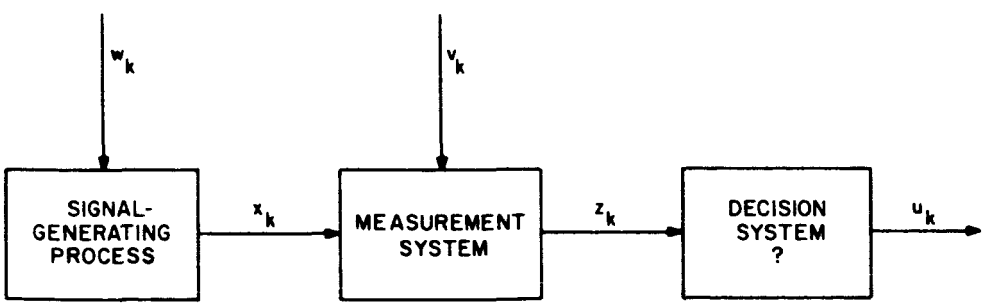

FIG. IV-1 OPTIMUM INFORMATION SYSTEM

Given (1) The known signal generating process

$$x_{k+1} = f(x_k, w_k, k) \qquad \text{(IV-1)}$$

(2) The measurement process

$$z_k = h(x_k, v_k, k) \qquad \text{(IV-2)}$$

(3) The probability density functions $p(x_0/\text{no measurement})$, $p(w_i)$ and $p(v_i)$, $i = 0, \ldots, N$, with w and v white.

Find The decisions u_k that minimize, on the average, a cost function of the form

$$J(u_0, \ldots, u_N) = E\{\sum_{k=0}^{N} l(x_k, u_k, k)\} \qquad \text{(IV-3)}$$

where the expectation is taken with respect to the random variable x_k.

It is shown in Appendix A that the solution to the stated problem is relatively simple if the decision u_k has no repercussion on future decisions u_{k+1} (*i.e.*, does not constrain such later decisions) and relatively complex in the alternative, where the mathematics of combined optimization theory apply.

For the purposes of this discussion, it suffices to consider the simple (yet meaningful) case of unconstrained decisions. Under these circumstances, the minimization of the functional (IV-3) reduces to the sequence of minimizations of the functions

$$J_k(u_k) = \underset{x_k}{E}\{l(x_k, u_k, k)\}$$

$$= \int_{x_k} l(x_k, u_k, k) p(x_k/Z_k) dx_k \quad . \qquad \text{(IV-4)}$$

The optimum decision u_k^* minimizes $J_k(u_k)$, that is

$$u_k^* = \arg \min_{u_k} J_k(u_k) \quad . \qquad \text{(IV-5)}$$

B. PROBLEM SOLUTION

It is readily seen from Eqs. (IV-4) and (IV-5) that the synthesis of the optimum decision system involves two steps, namely:

(1) The computation of the conditional probability density function $p(x_k/Z_k)$ of the state x_k given all available information, both prior and collected.

(2) The selection of the optimum decision u_k^* by minimization of $J_k(u_k)$.

The function $p(x_k/Z_k)$ can always be computed in principle by recursive application of Bayes's theorem, as discussed in Sec. III and in Appendix A. The determination of u_k^* follows from differentiation of $J_k(u_k)$ or, in the more general case, from a search over u_k.

C. MARSCHAK'S ILLUSTRATIVE EXAMPLE

In [Ref. 5], the example of an investor playing the stockmarket under simplifying assumptions is used throughout. This investor is allowed to reinvest his capital at every decision time k by buying the most promising stock, which he sells at a profit or a loss before the next decision time, $k + 1$. The optimum decision u_k^* maximizes the profit of the investor during the decision interval $[k, k + 1]$. The example is contrived, because the intelligent investor will try to maximize his profit over a decision interval $[0, N]$, which may be the duration of his life. Under those circumstances, the amount available for investment at $k + 1$ depends on the previous decisions.

In the example, the signal generating process is a mathematical model of the stockmarket, the imperfections of which are accounted for by w_k; the measurement process is the *Wall Street Journal* and the noisy information received from the stockbroker.

Even the simple problem formulation of unconstrained decisions is meaningful in the engineering sciences. As an obvious application, the problem of determining the best estimate of a state or a parameter (which is equivalent from the point of view of the mathematics) is cited. The quality of the estimate u_k of the actual state x_k is measured by

$$J_k = \underset{x_k}{E} \{l(x_k - u_k)\} \quad , \tag{IV-6}$$

where the function J_k is often of the weighted rms variety

$$J_k = \underset{x_k}{E} \{(x_k - u_k)^T Q (x_k - u_k)\} \tag{IV-7}$$

$$Q = \text{weighting matrix} \quad .$$

The most likly estimate $\hat{x}_{k/k}$ always minimizes Eq. (IV-7) regardless of the weighting matrix Q.

D. MEASURES OF INFORMATION

The conclusions obtained in Sec. II with regard to appropriate measures of information are illustrated and substantiated particularly well by the optimum information system formulation under discussion. These two measures are *quantity* of information and *value* of information.

From Eq. (IV-4), it follows that the conditional probability density function $p(x_k/Z_k)$ is necessary and sufficient to compute the cost of J_k. This function determines the quantity of information because it provides a complete summary of prior knowledge and collected data. A further important feature of $p(x_k/Z_k)$ is that it can be updated by means of a recursive relation of the form

$$p(x_{k+1}/Z_{k+1}) = \psi[p(x_k/Z_k),\ z_{k+1},\ k+1] \quad , \tag{IV-8}$$

which, in general, is not true for the successive moments such as mean and variance, the important exception being the linear Gaussian case.

The entropy H_k can neither be updated by a recursive relation of the form

$$H_{k+1} = \psi[H_k,\ z_{k+1},\ k+1] \quad , \tag{IV-9}$$

nor does it suffice to compute the cost J_k of Eq. (IV-4).

The elements of the covariance matrix $P_{k/k}$ of the function $p(x_k/Z_k)$ are useful in assessing the spread of the multivariate distribution and thus provide a measure of the quality of the information available; however, in the general case, these elements do not suffice to determine the cost J_k.

The value of the information received and available *a priori* is measured by the cost J_k, which is the main quantity of concern to the designer for the following two reasons:

(1) By expressing J_k in terms of parameters defining the measurement system (*e.g.*, the accuracy of a sensor), and the amount of prior knowledge (*e.g.*, the accuracy of the model), the critical sensor and model parameters can be pinpointed and cost trade-offs between alternate sensors can be found.

(2) By expressing J_k in terms of the parameters of a no decision system, the degradation of performance ent nonoptimal decision making can be assessed and the v optimal decision making established.

V OPTIMUM QUANTIZATION

The quantizer is an informtion handling element of frequent occurrence in control systems. It has the property that although its input, x, can take on any real value, its output, $q(x)$, must always be selected from some finite set. A quantizer is completely characterized by the set of output levels, denoted as c_0, c_1, ..., c_N, and the relation that determines which output level is to be used for each value of the input.

Customarily, this relation is determined by dividing the real line, $-\infty < x < \infty$, into $(N + 1)$ intervals such that each output level corresponds to one interval. These intervals are specified by the N switchpoints, , d_1, d_2, ..., d_N, where d_i is the value of x for which $q(x)$ changes from c_{i-1} to c_i. In general, the output levels are monotonically increasing with i, the index. The quantizer characteristic then appears as in Fig. V-1.

The work described in Appendix B provides a procedure for optimally designing a quantizer with a fixed number of levels, $N + 1$. The design procedure specifies the $(2N + 1)$ parameters of the quantizer, namely the $(N + 1)$ output levels and the N switchpoints. The criterion on which the design is based is the expected value of some function of the error between input and output of the quantizer. This criterion is a measure of how accurately the quantizer output, which must be chosen from a finite set, can approximate the input, which can have any real value. A fidelity criterion of this type maximizes the information transfer through the quantizer in the sense that it minimizes the average error introduced.

If $p(x)$ is the probability density function of the input, x, and if d_0 and d_{N+1} are set equal to $-\infty$ and to $+\infty$ respectively, then the criterion can be written compactly as

$$J = \sum_{i=0}^{N} \int_{d_i}^{d_{i+1}} g(x, c_i) p(x) dx \quad , \tag{V-1}$$

where $g(x, c_i)$ is the error measure when the input is x and the output is c_i. A typical error measure is squared error, in which case

$$g(x, c_i) = (x - c_i)^2 \quad . \tag{V-2}$$

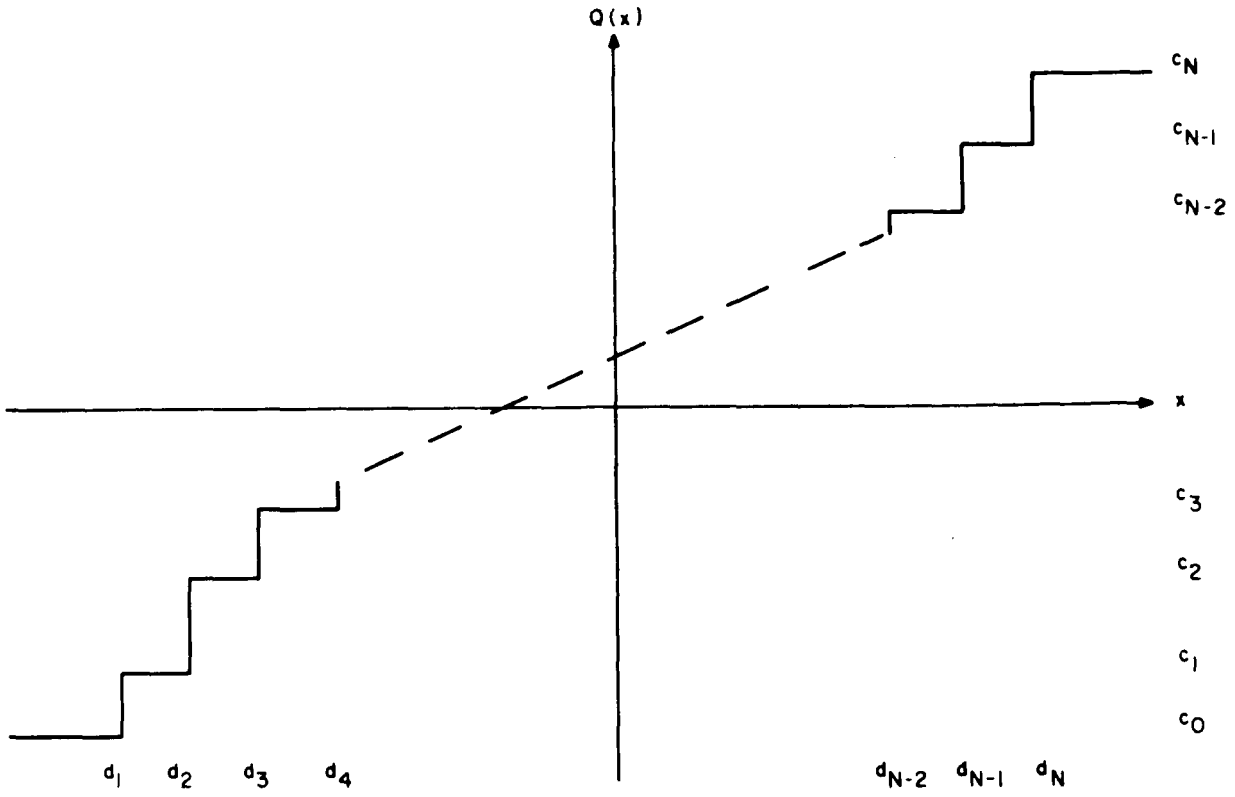

FIG. V-1 QUANTIZER CHARACTERISTICS

The design procedure consists of finding the $(2N + 1)$ values c_0, c_1, ..., c_N and d_1, ..., d_N that minimize J. This can be accomplished by differentiating J with respect to each of these $(2N + 1)$ parameters and setting the derivatives equal to zero. However, the design equations that result cannot be solved explicitly except in a very few cases. Nevertheless, one important result is obtained directly; under certain assumptions, which are met in almost all cases of practical interest, it can be shown that the switchpoints fall halfway between the output levels, that is

$$d_i = \frac{1}{2}(c_{i-1} + c_i) \quad . \tag{V-3}$$

This reduces the number of free parameters from $(2N + 1)$ to $(N + 1)$.

One way of solving the $(N + 1)$ remaining design equations is to search simultaneously over the $(N + 1)$ parameters until a set that satisfies the equations is found. However, this is not practical for large N.

In Appendix B, a procedure of much greater practical value is presented. The outstanding feature of this procedure is that it carries out the computations in such a way that a one-dimensional [rather than $(N + 1)$-dimensional] search suffices. As a result, it is possible to treat the case of large N. The procedure converges rapidly and it is ideally suited for implementation on a digital computer. It is described in detail in Sec. B of Appendix B. The procedure can easily be extended to handle many related problems. The case where the output of the quantizer is bounded is treated in the Appendix.

The procedure as described thus far is useful for static optimization because it depends only on the instantaneous values of the input and output of the quantizer. This point of view is appropriate when the quantizer is considered to be a part of an open-loop system, such as a measuring device. However, when the quantizer is introduced into the feedback loop of a dynamic system, the optimization problem becomes much more complex. In Sec. C of Appendix B some of the problems that arise are described. Early attempts[15,16] to treat this case resulted in techniques that are not computationally feasible for large N.

In Sec. D of Appendix B it is shown that for the case of linear plant equations, linear observations equations, Gaussian noise, and a quadratic performance criterion, the optimum quantizer design can be separated from the optimum design of the remainder of the system. A proof is presented for the case in which the quantizer is at the output of the controller, as in Fig. V-2. The extension to the case where the quantizer is located elsewhere in the feedback loop can easily be made.

The result that is obtained is that the overall optimum system design can be found by first designing the optimum feedback control system as if the quantizer were replaced by a unity gain and then using the previously discussed procedure to find the optimum quantizer characteristic. The fidelity criterion on which this latter optimization is based is the square of the error. Even when this procedure of separately designing the systems is not exactly optimum, a good approximation to the optimum system is often found. The simplification that results from the separation of the two designs is quite significant and the optimum or near-optimum synthesis of many practical systems containing quantizers thus becomes straightforward.

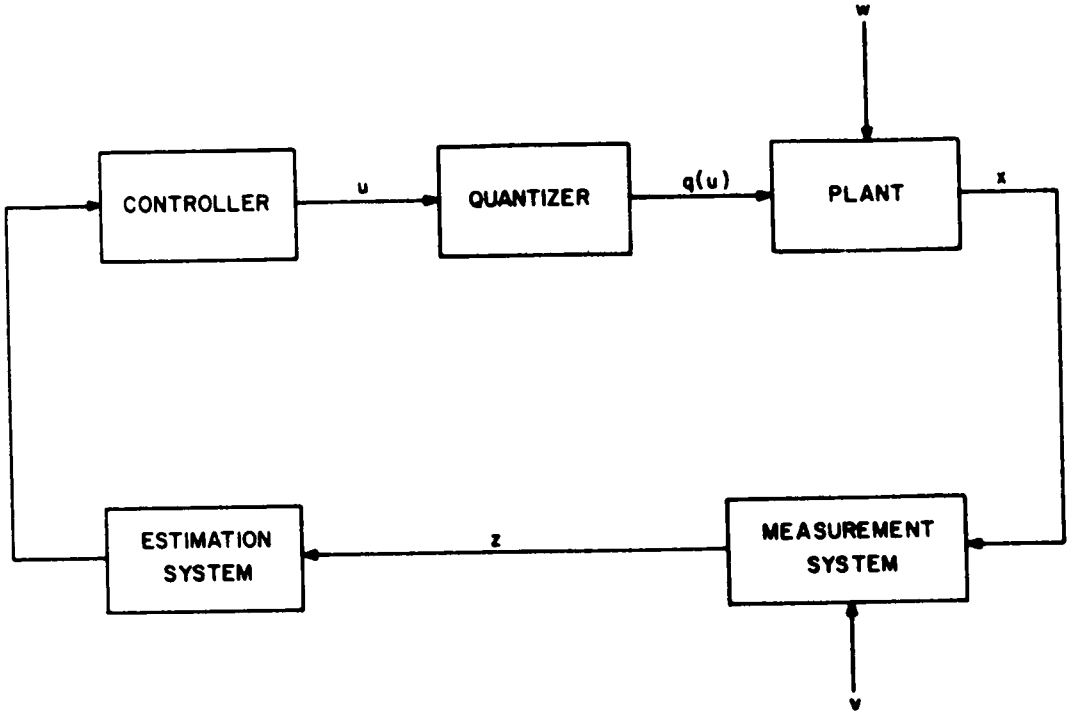

FIG. V-2 GENERAL DYNAMIC SYSTEM WITH QUANTIZED CONTROL

VI SYSTEMS WITH FIXED CONTROLLER

It is customary to design a control system for adequate transient and steady-state performance by neglecting, in the initial design phase, the random effects w and v. The resulting fixed controller relating the plant input u to the measurement system output z may or may not be altered thereafter to reduce the degrading effects upon performance of w and v.

The present section is concerned with the calculation of performance in the presence of such random perturbations and on the assumption that a fixed controller has already been laid down. The general nonlinear non-Gaussian case is considered and special cases, particularly those derived from well known methods of classical control theory, are treated.

A. GENERAL PROBLEM FORMULATION

For the closed-loop system shown in Fig. VI-1, the problem is formulated as follows:

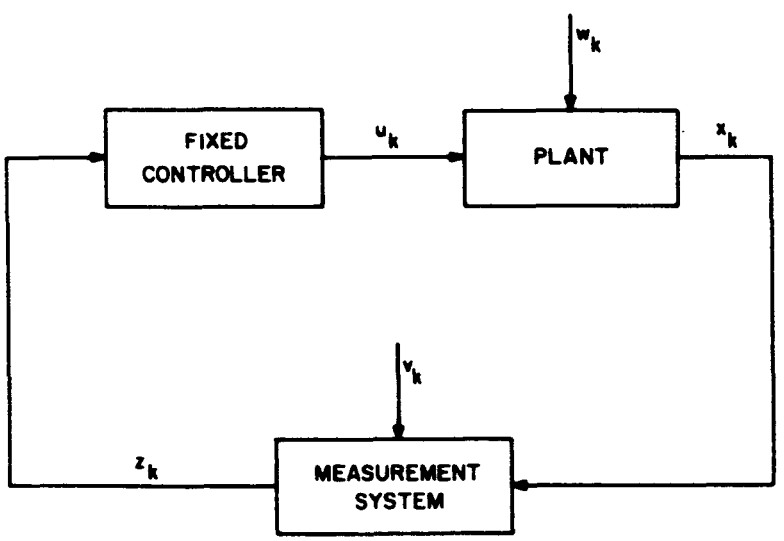

FIG. VI-1 SYSTEM WITH FIXED CONTROLLER

Given the plant, measurement and controller equations

$$x_{k+1} = f(x_k, w_k, u_k, k) \qquad \text{(VI-1)}$$

$$z_k = h(x_k, v_k, k) \tag{VI-2}$$

$$u_k = g(z_k, k) \tag{VI-3}$$

and the probability density function $p(x_0)$, $p(w_i)$, $p(v_i)$, w and v white $i = 0, \ldots, N$, find the expected cost

$$J = E\left\{\sum_{k=0}^{N} l(x_k, u_k, k)\right\}$$

for the interval $(0, N)$, the expectation being with respect to the random variables x_k and u_k.

The solution involves the two following steps:

Computation of the density functions $p(x_k)$ and $p(u_k)$—This computation can always be carried out, in principle, on a digital computer, but requires impractically large high-speed memories for even moderate dimensions of the state when the customary programming techniques are used.

It is shown in Appendix C that $p(x_k)$ is given by a stochastic difference equation of the form

$$p(x_{k+1}) = \psi[p(x_k), k] \tag{VI-4}$$

and that $p(u_k)$ can be computed in a straightforward manner once $p(x_k)$ is known.

Computation of the cost J—The computation of J proceeds without any formal difficulty, since the densities $p(x_k)$ and $p(u_k)$ are now known. Thus, the partial cost J_k incurred during the interval $[k, k+1]$ is simply

$$J_k = \int_{x_k}\int_{u_k} l(x_k, u_k, k)p(x_k)p(u_k)dx_k du_k \quad . \tag{VI-5}$$

The total cost J is the sum of all these partial costs

$$J = \sum_{k=0}^{N} J_k \tag{VI-6}$$

As stressed in previous section, $p(x_k)$ is necessary to determine the value of information, which is measured by J. An alternative method of calculating J by use of combined optimization theory is presented in Sec. III-D.

B. CLASSICAL SYSTEM PERFORMANCE MEASURES

In Appendix D, the relation between the information handling characteristics and performance of control systems is considered for classical performance measures such as pole location (for transient response) and steady-state error.

1. System Description

For this formulation, it is assumed that the plant, measurement system and controller are linear, and that the system operates in continuous time. Because of the linearity assumption, all components can be represented by appropriate transfer functions.

2. Sensor Imperfections

Some of the sensor imperfections that can be treated with this model are

(1) Constant biases

(2) Sensor gain changes

(3) Sensor dynamics

(4) Additive measurement noise

(5) Sampling (by use of z-transforms)

(6) Quantization (by replacement with an equivalent noise source).

3. Results

One objective of this study is to show that the classical methods of analysis based on Laplace transform theory can be applied to investigate the effect of the sensor defects listed above. Since these methods are well known, it was only necessary to illustrate how they can be applied in the case of a simple position servo. From this example, it is seen that classical control theory may be directly applied to an important class of problems involving imperfect sensors; the root locus approach, straightforward sensitivity considerations, and the use of power spectral density are of particular value.

In the second part of Appendix D, a method for designing systems to optimize classical criteria with a state space approach is presented. In addition a recently developed state space method for determining the sensitivity of pole locations to parameter changes is given. As in the first part, the ideas are illustrated with a simple position servo.

C. PROBABILISTIC FEEDBACK

Efforts to exploit the results of information theory in order to analyze information requirements of guidance and control systems have not been successful. A major stumbling block arises because the analytic property—the joint entropy of two independent events is the sum of the individual entropies, which is so useful in information theory—does not help much when considering control systems. In the latter we are much more often led to compute the entropy of the sum of two random variables, and it is not in general calculable from the individual entropies.

Useful results can be obtained if one considers certain marginal probability distributions. Specifically, let the system (plant and controller) have a state vector x of n elements obeying the continuous-time equations:

$$\dot{x} = Fx + Gw + Hy \qquad 0 \leq t < \infty \qquad \text{(VI-7)}$$

with matrices F, G, and H possibly time varying. Let $w(t)$ be a white random vector with ensemble mean zero and covariance matrix Q,

$$E(w) = 0, \quad E[w(t_1)w^T(t_2)] = Q\delta(t_1 - t_2) \quad .$$

Let $p_f(y|x)$ be the conditional probability density of the feedback term y; write the conditional mean m as

$$m = m(x) = \int dy\, y p_f(y|x) \qquad \text{(VI-8)}$$

and the ensemble conditional covariance as

$$E\{[y(t_1) - m_1][y(t_2) - m_2]^T | x(t_1), x(t_2)\} = S(x)\delta(t_1 - t_2)$$
$$= \delta(t_1 - t_2) \int dy\, [y - m][y - m]^T p_f(y|x) \qquad \text{(VI-9)}$$

so that y is probabilistically dependent on x and is white in time. Take y and w to be statistically independent, write $p(x,t)$ for the ensemble probability density of the system state at any time t, and take $p(x,0)$ to be given.

Then the evolution of the state $x(t)$ is a Markov process and the instantaneous state probability density $p(x,t)$ will obey the Fokker-Planck partial differential equation

$$\frac{\partial p}{\partial t} = -\sum_{k=1}^{n} \frac{\partial}{\partial x_k}(\alpha_k p) + \frac{1}{2}\sum_{k=1}^{n}\sum_{j=1}^{n} \frac{\partial^2}{\partial x_k \partial x_j}(\beta_{kj} p) \qquad \text{(VI-10)}$$

where

$$\alpha = Fx + Hm(x) \qquad \text{(VI-11)}$$

$$[\beta_{kj}] = GQG^T + HS(x)H^T \qquad \text{(VI-12)}$$

From this equation, by judicious use of integration by parts, one can obtain families of ordinary differential equations that will describe the time evolution of many quantities defined by integrals of functions of x over the $p(x,t)$ density. Complete families of equations can be found, for example, for the instantaneous ensemble moments of the state in any linear problem and also in a number of interesting problems involving nonlinear probabilistic feedback (*e.g.*, where the feedback "noise" is dependent on the state).

This problem formulation and the resulting Fokker-Planck equation also apply to the case in which the feedback term y is *discretized*, so we can treat the system with a quantizer and digital communication channel in the feedback path. It turns out that some approximations are needed in order to obtain a complete set of ordinary differential equations for the state mean and covariance matrix; the first approximations that were tried introduced a bias in the steady-state results. However, even these should give good answers when the number of quantization levels is fairly large (say, more than eight). These results thus permit the designer to analyze the response of his control system design in terms of time histories of its ensemble means, variances, etc.; the effects of both system dynamics and information-handling-element characteristics (inaccuracies, averaging, communication channel errors, etc.) on system response can be examined in detail.

D. APPROXIMATE DESIGN OF A FIXED CONTROLLER SYSTEM WITH DIGITAL FEEDBACK PATH

1. Introduction

To illustrate the power of straightforward engineering arguments and sound approximations, the example problem shown in Fig. VI-2 is considered.

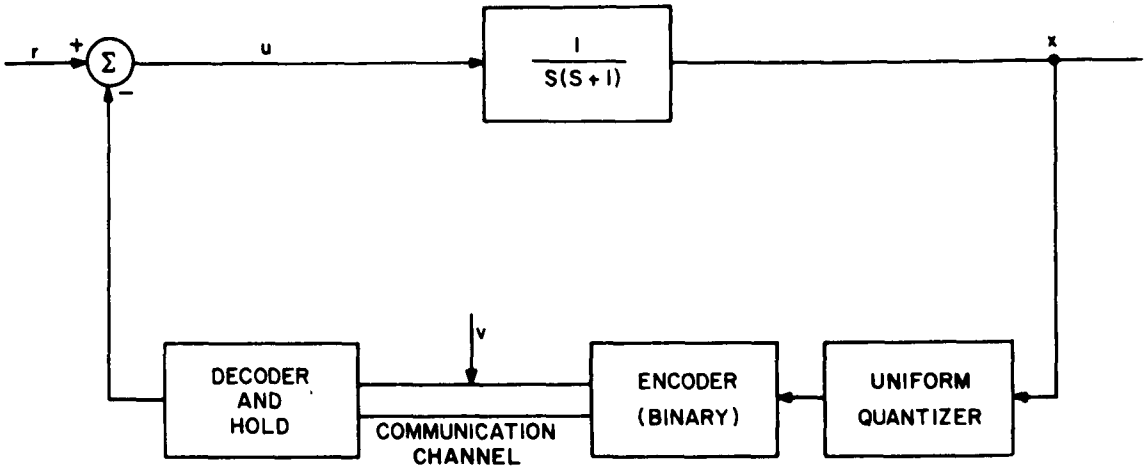

FIG. VI-2 SYSTEM WITH NOISY DIGITAL FEEDBACK

The quantizer has N equally spaced levels, which are converted into an n-bit message in the encoder. This message is sent over a frequency shift keying (FSK) communication channel, where each bit is changed with a probability p_v in accordance with

$$p_v = \frac{1}{2} e^{-\alpha \delta t} \qquad \text{(VI-13)}$$

where

α = constant characterizing the quality of the channel

δt = time interval for transmitting one bit.

The digital message received with a transmission delay Δt is converted into an analog signal in the decoder and compared to the command input r to generate a correcting signal u.

The design parameters p_v, N, n, δt and Δt are related by (VI-13) and the additional two equations

$$N = 2^n \tag{VI-14}$$

$$\Delta t = n\delta t \tag{VI-15}$$

with the result that (VI-13) can be rewritten as

$$P_v = \frac{1}{2} l^{-\alpha \frac{\Delta t}{n}} \tag{VI-16}$$

For convenience, the bit error probability p_v and the message time Δt are retained as the free design parameters and the problem consists of minimizing the steady-state rms output $E\{(r - x)^2\}$ by selecting optimum values of p_v and Δt.

To illustrate the situation, the following two extreme cases are considered:

(1) Δt is small compared to the system time constant and the error $(r - x)$ is affected by nonzero mean noise, the major effect of which is the introduction of a bias; the magnitude of this bias depends on the transmitted code word, *i.e.*, on the output x. The amplitude of the transients resulting from an erroneous message are negligible because Δt, the time during which the resulting erroneous control prevails, is much smaller than the system's dominant time-constant. This situation will be referred to as the *steady-state* mode of operation.

(2) Δt is large compared to the system's dominant time constant; transmission errors are now relatively infrequent, but set up a non-negligible transient since they cause an erroneous control signal, which subsists for a relatively long time-interval. In addition to this infrequent transient error, there is a deadband error resulting from the fact that no correcting signal is generated as long as the output remains in the deadband of the quantizer. This situation will be referred to as the *transient* mode of operation.

2. Separation Between Steady-State and Transient Modes

The average frequency with which an erroneous message is received is

$$f_a = \frac{p_v n}{\Delta t}$$

whereas the break-frequency of the linearized system of Fig. VI-2 is

$$f_b = \frac{1}{2\pi} \quad .$$

The steady-state mode prevails for

$$f_a > f_b$$

and the transient mode for

$$f_a < f_b \quad ,$$

with a separation given by

$$\frac{p_v n}{\Delta t} = \frac{1}{2\pi} \quad . \tag{VI-17}$$

With Eqs. (VI-14) and (VI-15), this separation is expressed in terms of the single free design parameter p_v as

$$\log 2p_v + 2\pi\alpha p_v = 0 \quad . \tag{VI-18}$$

3. Performance in the Steady-State Mode

The performance is measured by the average square of the output offset resulting from the bias effects of the noisy communication channel, the average being taken over time and the space of the input commands r.

It is proved in Appendix F that the number j of quantization level offsets caused by this bias is approximately related to r as shown in Fig. VI-3.

For r uniformly distributed in the range

$$-\frac{X}{2} \leq r \leq \frac{X}{2} \tag{VI-19}$$

the resulting cost J_1 is

$$J_1 = \underset{r}{E}\{(r - x)^2\} = \frac{N^2 p_v 2}{3} \quad . \tag{VI-20}$$

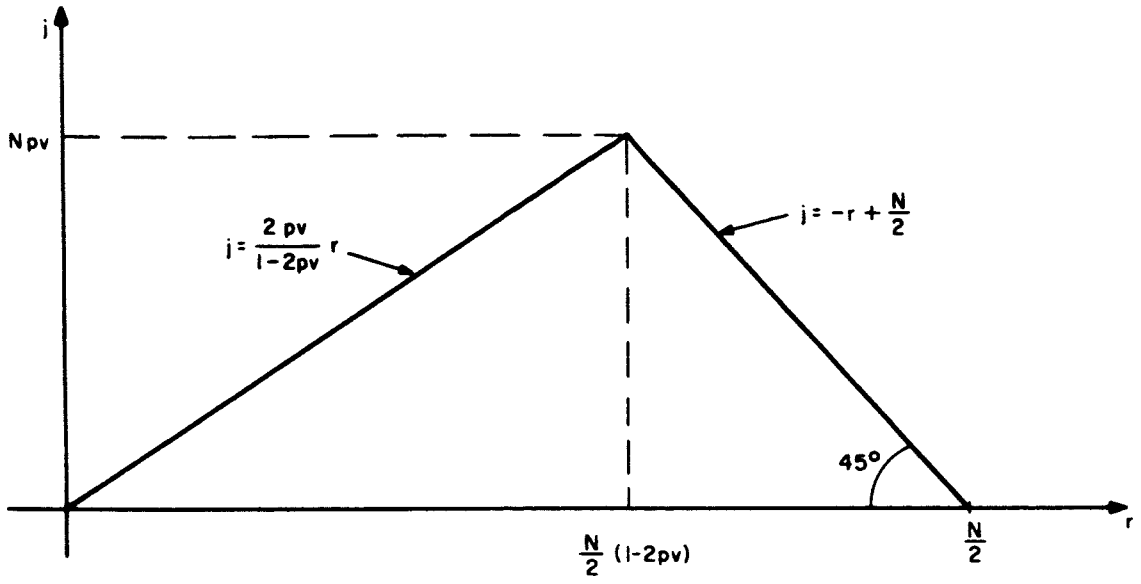

FIG. VI-3 QUANTIZATION LEVEL OFFSETS, j, IN TERMS OF r, p_v, AND N, FOR A BINARY CODE

4. PERFORMANCE IN THE TRANSIENT MODE

In this mode, it is assumed that the following two errors occur:

(1) The deadband error, which is assumed to be uniformly distributed within the deadband determined by the quantizer step spacing X/N. The resulting cost J_2' is

$$J_2' = E\{(r - x)^2\} = \frac{X^2}{3N^2} \quad . \tag{VI-21}$$

(2) The transient error, which is caused by an erroneous control applied during the interval Δt when an erroneous message is received. Denoting by ρ the magnitude of the channel error at the decoder output and by $\mathcal{J}(t,\Delta t)$, the unit impulse response of the linearized system of Fig. VI-2 operating with the channel delay Δt the integral squared error is

$$\text{ISE} = \rho^2(\Delta t)^2 \int_0^\infty \mathcal{J}^2(t,\Delta t)dt \quad . \tag{VI-22}$$

The frequency of occurrence of this error is determined by p_v and the magnitude of its effect is measured by a function C of the number of quantization increments X/N and the particular code chosen.

The cost associated with this second error is averaged over time with the result that

$$J_2'' = (\Delta t)^2 p_v C \int_0^\infty \mathcal{J}^2(t,\Delta t)dt \quad . \tag{VI-23}$$

It is shown in Appendix F that $\mathcal{J}(t,\Delta t)$ is fairly insensitive to Δt for Δt sufficiently small ($\Delta t < 0.3$ sec) and that C is a constant for N sufficiently large ($N > 8$) with the result that (VI-23) can be evaluated very easily.

The costs J_2' and J_2'' are added to provide the total operating cost J_2 for the transient mode.

5. Design Chart

With expressions for the costs J_1 and J_2, it is now easy to establish a design chart suggesting optimum values of the free design parameters p_v and Δt. This is done in Fig. VI-4 for a communication channel characterized by $\alpha = 63.4$ and a quantizer range $X = 1.5$.

It is seen that the best design parameter selection is approximately

$$\Delta t = 0.5$$

$$p_v = 10^{-3}$$

$$n = 5$$

$$N = 32 \quad .$$

Larger values of Δt are ruled out, since they lead to poor transient response and cause instability for $\Delta t > 0.8$.

The separation between steady-state and transient modes is at $p_v = 10^{-2}$; consequently, the best parameter selection corresponds to the transient mode. The costs J_2' and J_2'' are of the same order of magnitude.

It is clear from Fig. VI-4 that the costs can become extremely high for poor choices of design parameter.

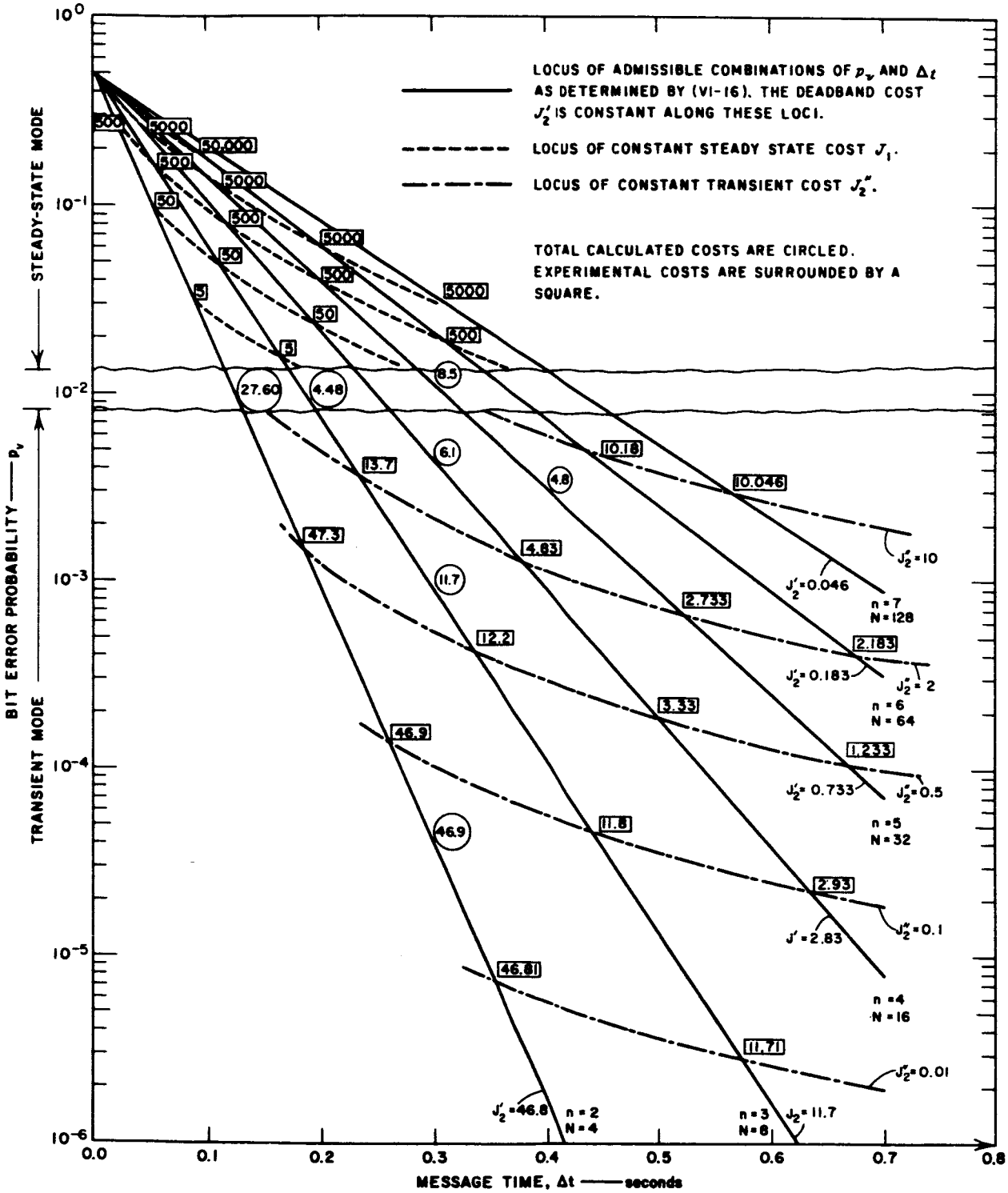

FIG. VI-4 DESIGN CHART FOR SYSTEM WITH NOISY DIGITAL FEEDBACK PATH

6. Experimental Verification

In order to check the validity of the numerous simplifying assumptions made, a Monte Carlo simulation program was written. The numerical checkpoints provided by this simulation are shown in squares on Fig. VI-4. It is seen that the assumptions made are well justified.

7. Information-Theoretic Implications

The channel capacity of a symmetric binary channel is expressed as

$$C_1 = ln\ 2 + p\ ln\ p + (1 - p)\ ln\ (1 - p)$$

where C_1 is capacity in nepits per bit and p is the bit error probability. The capacity in nepits per second is obviously

$$C = C_1 R \quad ,$$

where R is the rate in bits per second.

For the FSK system with incoherent detection the bit error probability is given by

$$p = \frac{1}{2} e^{-\alpha \delta t}$$

where α is a power parameter (signal power divided by one-sided noise spectral density) and δt is the bit time. Then

$$ln\ 2p = -\alpha \delta t = -\frac{\alpha}{R}$$

and

$$C = [ln\ 2 + p\ ln\ p + (1 - p)\ ln\ (1 - p)] \cdot \frac{\alpha}{-ln\ 2p}$$

Thus C/α is a function of p alone and some values are given in Table VI-1. It is noted that there is a fairly broad maximum running from about $p = 0.065$ to $p = 0.165$ with the absolute maximum at $p = 0.110$.

Note that the location of the maximum is independent of α and occurs at probabilities above the boundary between "transient" and "steady-state" system behaviour. Since other considerations indicate that the system should be operated at lower values of p, the communication channel will not be operated close to its maximum capacity. Note also from Fig. VI-4 that knowledge of the channel capacity alone gives no clue concerning the performance of the system.

Table VI-1

VARIATION OF C/α WITH p

p	C/α
0.005	0.144
0.045	0.212
0.065	0.222
0.090	0.228
0.110	0.229
0.130	0.228
0.165	0.221
0.200	0.210
0.300	0.161

VII SUMMARY AND CONCLUSIONS

The chief objective of the study was to relate the performance of control and guidance systems to the information-handling characteristics of their key constituents. To accomplish this objective, the following questions were studied:

(1) What is information in this context and what are quantitative descriptions of information?

(2) Given a control system with fixed controller, how does one relate performance to the system parameters?

(3) Given a control system with fixed plant and measurement subsystem, how does one design a controller that optimally utilizes the information available *a priori* and collected in actual operation?

From the answer to these questions, the desired relation between performance and the information-handling characteristics of the subsystems (measurement as well as controller subsystems) follows directly in principle, although the actual calculations may exceed the capabilities of present day computers in many practically important situations. These relations may become very complex, as evidenced, for instance, by the concept of dual control; one must recognize that a complex problem usually leads to a complex answer.

A. MEASURES OF INFORMATION

In order to calculate the performance of a control system affected by random forces, notably plant perturbations w and measurement noise v, the following two mathematical notions are required in the general case:

(1) The condition of the system must be expressed un terms of its state, which summarizes the complete past history of the system. Description by Laplace or z-transforms is possible for linear stationary systems with Gaussian noise and fixed controller, but out of the question for all other cases.

(2) The effects of the random forces must be expressed in terms of probability density functions. In special cases, these probability density functions may be replaced by a finite number of moments, which constitute a sufficient statistic.

With these descriptions, it is then possible to compute system performance for the case of a fixed controller and the case of an optimum controller. The quantity required to compute performance is the probability density of the state, conditioned on *a priori* information only in case of the fixed controller and conditioned on all available information in the case of optimum controller. This probability density function is determined, among other effects by the random effects w and v. It constitutes a description of the uncertainty about the state in terms of the statistics of w and v, and hence can be used to determine the *quantity* of information available about the state. Although the probability density is not the quantity of chief concern, it is required for the calculation of the system performance J which is the quantity of chief concern. The scalar J determines the *value* of the information provided not only by the measurement system, but also available *a priori*; it furthermore measures the efficiency with which this information is processed in the controller. Once J has been calculated, the sensitivity of J with respect to measurement, processing, and control subsystem parameters can be established, either analytically as for the linear Gaussian case, or by machine computation for the general case.

The performance J also establishes the need for feedback in terms of the performance requirements, the random perturbations w, and the initial uncertainty about the state. Specifically, it tells under what circumstances closed-loop control is necessary to meet the performance requirements.

One important conclusion which results from these considerations is that the classical information theory is of no direct help to the designer of guidance and control systems, except under very special circumstances, such as Kelly's gambler[5,8]. On the other hand, control theory is of considerable assistance to the designer of information systems as evidenced by Sec. IV on the design of optimum information systems.

B. SYSTEMS WITH FIXED CONTROLLER

The design of most control systems—notably simple systems with a small number of states, such as position control servos—usually consists

of selecting a fixed controller structure (*e.g.*, position and rate feedback) to provide acceptable transient and steady-state performance in the absence of random perturbations. After a basic design has been obtained, the effects of random forces and measurement system deficiencies are thereafter determined and controller adjustments and filters are used to account for these effects.

For fixed controller design, it is always possible and frequently straightforward to relate performance to the measurement system characteristics, the controller parameters and the environmental perturbations w. The following four approaches were taken:

(1) Measurement system deficiencies were characterized by bias errors, additive noise, internal dynamics and nonlinearities such as quantization. For linear systems, Caussian noise and quadratic performance measure, the relation between performance and these deficiencies can be derived by well-known results of classical control theory. The effects of quantization and other nonlinearities can be assessed approximately. This work, which is of direct practical usefulness to the control engineer, is described in detail in Appendix D.

(2) In the general case of nonlinear systems, non-Gaussian random effects, and non-quadratic performance measure, a closed form relation between performance and the relevant system parameters cannot be obtained. But it is always possible in principle, and this does not seem to be a generally known fact, to obtain this relation by machine computation. This computation involves two steps:

   (a) Determination of the probability density function of the state $p(x_k)$.

   (b) Calculation of performance J on the basis of $p(x_k)$.

In practice, these computations become laborious, even for systems with a moderately high number of states, when the customary programming procedures are used.

The computational approach for this general case is spelled out in detail in Appendix C.

(3) For the less general case of a linear plant and controller and a nonlinear measurement system with non-Gaussian noise, the Fokker-Planck equation, a partial differential equation whose solution is the probability density function of state with time, was derived and approximated by a set of differential equations giving mean and variance with time. The inadequacy of entropy as well as the difficulty of actually computing it recursively was pointed out in this study, of which a detailed account is given in Appendix E.

(4) To illustrate the potential of approximate design techniques, the case of a linear system with a noisy digital communication channel in the feedback path was considered. Performance was calculated under very simple approximations and the results were checked by means of a Monte Carlo simulation. The accuracy of these results are quite adequate for many simple systems. A by-product of this study, discussed in detail in Appendix F, is the proof by example that there exists no strong relation between system performance and channel capacity.

C. SYSTEMS WITH CONTROLLERS MAXIMIZING THE UTILIZATION OF INFORMATION

In order to relate performance to the information-handling characteristics assuming that optimum utilization of information is made, the combined optimization problem was formulated and a solution was derived with dynamic programming. It was shown that the control sequence leading to optimum performance is determined by the conditional probability density function $p(x_k/Z_k)$, where Z_k incompasses all *a priori* and collected information. This is, further, an illustration of the fact that the quantity of pertinent information is a property of the probability density of state.

The combined optimization problem has a feature which is uncommon to the more conventional control problems, but very common to human decision processes, namely dual control: as long as information is insufficient, the actions which provide most information are taken instead of actions which are directly aimed at achieving the control objectives. This procedure permits better satisfaction of the ultimate objective at a later time.

In the general case, the dual control problem does not have a practically computable exact solution. However, important special cases which can be resolved with present day machines were identified, notably:

1. The Linear Gaussian Case with Quadratic Performance

In this case, the control u_k does not affect the conditional covariance of the state; because control action cannot be used to gain information, the dual control aspect hence does not enter.

For the linear case, concise mathematical expressions were developed to relate performance to the characteristics of the measurement system and the amount of prior information. Specifically, the effect upon performance of the lack of information is determined by the conditional

covariance of the state. The value of more information, *i.e.*, the possible increase in system performance, depends on the change in this conditional covariance as well as on other terms related to the objective function and the plant. With the help of these mathematical expressions, the designer can easily determine which state-variables must be determined accurately (either by direct measurement or by computation from other measured state variables) and which can be neglected. Also, the degrading effect of sub-optimal processing, including neglect of prior information about the initial state, can be assessed.

2. Optimum Information Systems

In this special case, decisions made on the basis of observations do not affect the dynamic process generating these observations, *i.e.*, the loop is not directly closed (it may be closed indirectly because of constraints). This case has practical importance in a variety of situations ranging from the interpretation of test data to optimum strategies for playing the stock-market, the latter being Marschak's example problem, which prompted the study of information systems.

Much related work is available in the literature in the field of decision theory. This study, however, adds several novel elements, *i.e.*,

- inclusion of the dynamics model governing the signal generating process to utilize past observations
- consideration of dynamic, as opposed to the customary static, performance measures
- discussion of the relations between performance and information. These optimum information systems provide a particularly good illustration of the required mathematical description of information—the conditional probability density function $p(x_k/Z_k)$—and of the value of information.

A major effort was devoted to combined optimization theory, which is believed to be as important as the theory of optimum control and the theory of optimum estimation, not only in terms of its immediate practical applications, but also in terms of its extensions to important systems concepts not specifically considered in this study. Thus, combined optimization theory provides the required mathematical framework for "adaptive" and "learning" systems where either the plant parameters or the statistics of the perturbations or both are not accurately known initially. Similarly, it leads to the optimum design of systems where the measurement subsystems

as well as the plant can be controlled. Additional important extensions relating to the theory of optimum classification and the theory of differential games were pointed out.

3. The Optimum Design of Systems Containing Quantizers

Prior to the effort in combined optimization theory, a study involving systems containing quantizers was carried out. Although this study did not shed much light upon the relations between performance and information, it generated a very useful approach to practically design such systems for optimum and near-optimum performance. The main elements of this approach are:

(1) Under certain commonly made restrictive assumptions, optimum system performance results if the controller, estimator and quantizer subsystems are optimized separately.

(2) The optimization of the quantizer (that is, the selection of the optimum step sizes and switchpoints for a given quantizer input probability density function) is performed easily by means of an efficient computational scheme developed in the course of the study.

D. PRACTICAL IMPLICATIONS OF COMBINED OPTIMIZATION THEORY

Assuming that the computational requirements can be overcome by means of rational approximations and efficient processing of the data, combined optimization is directly applicable to the design, evaluation and real-time control of dynamic systems affected by measurement noise and described by imperfect models.

Design—By providing the designer with relations between performance and the characteristics of the various subsystems, best choices in terms of performance improvement *vs.* dollar cost of these subsystems can be made.

Evaluation—The optimum solution provided by combined optimization theory sets a standard of comparison and thus indicates how well a given system performs and where significant improvements can be obtained.

Real-Time Control—With the results of combined optimization theory, real-time systems for which optimum control in the presence of noise measurement and imperfect models is important, can be synthesized.

E. PRACTICAL DIFFICULTIES OF COMBINED OPTIMIZATION THEORY

As is the case with most control system optimization procedures the computational requirements quickly exceed the capabilities of present-day machines for both computation time and high-speed memory unless special conditions hold.

It should be pointed out that these difficulties stem from the nature of the control problem rather than the method of solution. For feedback control, it is necessary to derive the control law as a function of the state if it is known or the conditional probability of the state if the state is not known. In either case, for high-dimensional systems, one is faced with the necessity of computing and storing a function of a large number of variables. Only in special cases where open loop control is suitable can such procedures as the gradient method greatly reduce computational requirements.

These computability problems have become so frequent—for an entirely different example, see Appendix C on the recursive calculation of probability density functions—that attention must be devoted to approximations and computational shortcuts. In the case of combined optimization, there exist several logical approaches to both problems, such as linearization and state increment dynamic programming, but they are by no means the only approaches and certainly not always the best, depending on the special features of specific problems.

ACKNOWLEDGMENT

The authors wish to acknowledge the guidance and assistance received from Messrs. B. Doolin, E. Steward, G. Smith, and R. Peery of the NASA Ames Research Center in their selection of the study and the definition of its objective.

REFERENCES

1. J. Peschon, W. H. Foy, and L. Meier, "Information Requirements for Guidance and Control Systems," Interim Technical Report, Contract NAS 2-2457, SRI Project 5237, Stanford Research Institute, Menlo Park, California (May 1965).

2. L. Meier, "Combined Optimal Control and Estimation Theory," NASA CR-416, 1966.

3. R. E. Bellman, *Adaptive Control Processes*, Chapt. 10, pp. 152-159 (Princeton University Press, Princeton, New Jersey, 1961).

4. Yu Chi Ho and R. C. K. Lee, "A Bayesian Approach to Problems in Stochastic Estimation and Control," 1964 JACC, Stanford University.

5. J. Marschak, "Remarks on the Economics of Information," Cowles Foundation Discussion Paper No. 70.

6. J. G. Truxal, *Automatic Feedback Control System Synthesis*, Chapt. 1 (McGraw-Hill Book Co., Inc., New York, New York, 1955).

7. C. E. Shannon, "The Mathematical Theory of Communication," *Bell System Tech. J.* 27, Nos. 3 and 4, pp. 379-423 and 623-656 (July 1948 and October 1948); also published as Claude E. Shannon and Warren Weaver, *Mathematical Theory of Communication* (University of Illinois Press, Urbana, Illinois, 1949).

8. J. L. Kelly, Jr., "A New Interpretation of Information Rate," *Bell System Tech. J.* 35, No. 4, pp. 917-926 (July 1956).

9. R. E. Kalman, "On the General Theory of Control Systems," *Proc. 1st Internat. Conf. Automatic Control* (Butterworth, London, 1962).

10. W. M. Wonham, "Stochastic Problems in Optimal Control," *IEEE National Convention Record* (1963).

11. E. Parzen, *Modern Probability Theory and Its Applications* (John Wiley and Sons, Inc., New York, New York, 1960).

12. R. E. Kalman and J. S. Englar, "An Automatic Synthesis Program for Optimal Filters and Control Systems Program B," Final Report, Contract NAS 2-1107, RIAS Division of the Martin Company, Baltimore, Maryland, (July 1964).

13. A. A. Feldbaum, "Dual Control Theory, Part I," *Automation and Remote Control* 21, No. 9, pp. 874-880 (September 1960).

14. A. A. Feldbaum, "Optimal Systems," Chapt. 7 in *Discipline and Techniques of Control Systems*, Ed. John Peschon (Blaisdell Publishing Company, New York, New York, 1965).

15. J. F. Tou, *Optimum Design of Digital Control Systems* (Academic Press, Inc., New York, New York, 1963).

16. J. Peschon, R. E. Larson, and A. S. Chen, "Optimum Design of Quantized Control System," Memo No. 2, Contract NAS 2-2457, SRI Project 5237, Stanford Research Institute, Menlo Park, California (February 1965).

APPENDIX A

OPTIMUM INFORMATION SYSTEMS

The purpose of the present appendix is to discuss and expand J. Marschak's paper "Remarks on the Economics of Information," Cowles Foundation discussion paper #70, one of the few references where meaningful mathematical definitions of information are given.

In what follows, Marschak's problem formulation is expanded to include a priori knowledge about the signal generating process and to provide the required recursive relations for the optimum processing of the information received at the output of a noisy communication channel or measurement system. These recursive relations are derived from Bayes's rule,* and constitute a generalization of the Kalman-Bucy estimator.

The problem formulation is much simpler than the combined optimization problem[2] in the sense that the decisions made at the output terminals of the measurement system do not affect the signal generating process. This formulation has great practical importance for the selection and design of complex measurement systems such as required for space exploration. The formulation has the further merit of leading to computable solutions in a large number of applications without the need for approximations.

In addition to these practical considerations, the problem formulation provides an excellent basis for discussing the potential and the shortcomings of information theory, and for deriving other and much more meaningful measures of information than entropy.

A. Problem Formulation

With reference to Fig. A-1, the problem is formulated as follows:

---

*References are listed at the end of each Appendix.

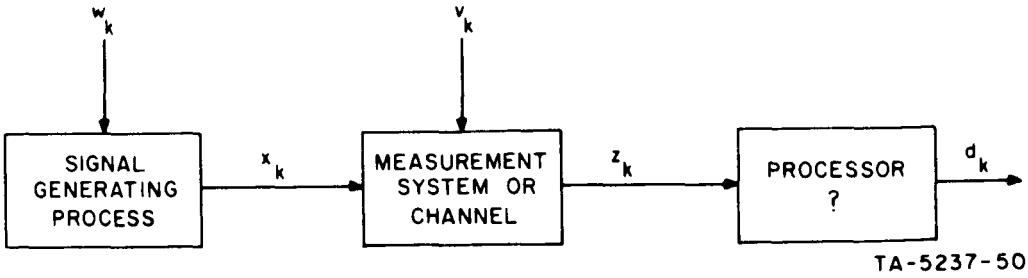

FIG. A-1 OPTIMUM INFORMATION SYSTEM

x_k = state at discrete time k of the signal generating process

z_k = measurement vector at time k

d_k = decision at k

w_k = random perturbation affecting the signal generating process

v_k = random noise affecting the measurement system.

Given:

1. The known signal generating process

$$x_{k+1} = f(x_k, w_k, k) \tag{1}$$

2. The intial probability density function (p.d.f.)

$$p(x_0/\text{no measurement})$$

3. The known equations governing the measurement system

$$z_k = h(x_k, v_k, k) \tag{2}$$

4. The p.d.f. of w_k, supposed white

$$p(w_k)$$

5. The p.d.f. of v_k, supposed white

$$p(v_k)$$

6. A cost function J of the actual state x and the decisions d to be minimized on the average.

Find:

A processor which minimizes J given all the a priori information and the measurements z_k up to and including k.

In the simplest case, which is also the one treated by Marschak, a cost J_k is incurred at every decision time k, and the decision d_k has no repercussion on future decisions. Thus

$$J_k(d_k) = \underset{x_k}{E} \{ \ell(x_k, d_k, k) \} \tag{3}$$

where the symbol $\underset{x_k}{E}$ denotes the expectation over the random variable x_k, that is

$$J_k(d_k) = \int_{x_k} \ell(x_k, d_k, k) \; p(x_k/Z_k) dx_k \tag{4}$$

where

$$Z_k \triangleq [z_0, \ldots, z_k]^T$$

The optimum decision d_k^* is clearly that admissible value of d_k which minimizes $J_k(d_k)$,

$$d_k^* = \arg \min J_k(d_k)$$

From this simple optimum decision process, the following two important conclusions are derived concerning the tasks which the processor must accomplish:

1. It must obtain the p.d.f. $p(x_k/Z_k)$, i.e. the probability of x_k conditioned on all available information, both a priori and measured.
2. It must compute J_k and find d_k^*.

B. Calculation of $p(x_k/Z_k)$

As pointed out in Refs. 2 and 3, the best approach to compute this conditioned p.d.f. is Bayes's theorem, which is first recalled.

Given two dependent random variables A and B, Bayes's theorem states that

$$p(A/B) = \frac{p(B/A)\ p(A)}{p(B)} \tag{5}$$

Or, returning to the stated problem

$$p(x_k/Z_k) = p(x_k/z_k,\ z_{k-1},\ldots,z_0)$$

$$= \frac{p(z_k/x_k, z_{k-1},\ldots,z_0)\ p(x_k/z_{k-1},\ldots,z_0)}{p(z_k/z_{k-1},\ldots,z_0)}$$

$$= \frac{p(z_k/x_k)\ \ p(x_k/Z_{k-1})}{p(z_k/Z_{k-1})} \tag{6}$$

The three p.d.f.'s on the right side of (6) are now considered separately:

1. The term $p(z_k/x_k)$ depends solely on the quality of the measurement system and can always be calculated in terms of $p(v_k)$ by using the measurement equation (2).

   From the chain rule

   $$p(z_k/x_k) = \int_{v_k} p(z_k/x_k,\ v_k)\ \ p(v_k)\ dv_k$$

   where z_k assumes the value specified by (2) with probability one for given x_k and v_k.

2. The term $p(x_k/Z_{k-1})$ which depends on the signal generating process only is further broken down by the chain rule as follows:

$$p(x_k/Z_{k-1}) = \int_{x_{k-1}} p(x_k/x_{k-1})\ p(x_{k-1}/Z_{k-1})\ d\,x_{k-1} \qquad (7)$$

The term $p(x_k/x_{k-1})$ can always be computed in terms of $p(w_{k-1})$ from equation (1). The term $p(x_{k-1}/Z_{k-1})$ results from the computation at k-1.

3. The term $p(z_k/Z_{k-1})$ only plays the role of a normalizing factor to ensure that $p(x_k/Z_k)$ integrates out to one. It is computed easily by integrating the two numerator terms

$$p(z_k/Z_{k-1}) = \int_{x_k} p(z_k/x_k)\ \ p(x_k/Z_{k-1})\ d\,x_k \qquad (8)$$

C. The Cost Function J

Generally speaking, it will be assumed that the cost function will depend on the state x and the decision d as indicated by the variational expression

$$J(D_N) = \underset{X_N}{E}\left\{\sum_{k=0}^{N} \ell(x_k,\ d_k,\ k)\right\} \qquad (9)$$

where

$$D_N = [d_0, \ldots, d_N]^T$$

$$X_N = [x_0, \ldots, x_N]^T$$

and where the decision interval [O,N] may be finite or infinite.

Three distinct cases are now considered, viz:

1. Any decision d_k made at k has no repercussion on later decisions d_{k+i}. This is the problem treated previously.

2. All observations are made before the decision process commences. If present decisions have a repercussion on future decisions, the resulting problem is a stochastic optimum control problem; if not, it is a special case of 1.

3. Both observations and decisions are made at each time k, and present decisions constrain future decisions. This is a combined optimization problem.

Case 1

Since present decisions do not constrain future decisions, minimization of the functional $J(D_N)$--see Eq. (9)--reduces to a minimization, at each time k, over the function

$$J(d_k) = \underset{x_k}{E} \{\ell(x_k, d_k, k)\} \tag{10}$$

which is the case previously treated.

Case 2

A repercussion of d_k upon future decisions d_{k+i} is mathematically accounted for a difference equation of the form

$$x'_{k+1} = \varphi(x'_k, d_k, k) \tag{11}$$

where the state x'_{k+1} governing the decision process must be admissible

$$x'_{k+1} \in X'$$

Let z_0 be the last observation and let $[0,N]$ be the interval over which the decisions are made, i.e.,

$$J = E\left\{ \sum_{k=0}^{N} \ell\,(x_k, d_k, k) \right\} \tag{12}$$

subject to

$$x'_{k+1} = \varphi(x'_k, d_k, k) \tag{11}$$

This is the formulation of the stochastic optimum control problem, given by Bellman with x_k playing the role of a random perturbation with known statistics $p(x_k/Z_0)$.* This conditional probability density function is precomputed recursively from

$$p(x_k\ Z_0) = \int_{x_{k-1}} p(x_k\ x_{k-1})\ p(x_{k-1}/Z_o)\ d\ x_{k-1} \tag{13}$$

starting at $p(x_1/Z_0)$ and using (1) to evaluate $p(x_k/x_{k-1})$.

A solution to the stochastic optimization problem is provided by dynamic programming, the appropriate search algorithm being

$$I(x'_k, k) = \min_{d_k} E_{x_k} \{ \ell(x_k, d_k, k) + I[\varphi\,(x'_k, d_k, k), k] \} \tag{14}$$

where $I(x'_k, k)$ is the minimum cost in the interval $[k,N]$ for the initial decision process state x'_k.

Case 3

When the decision process is governed by (11) and observations continue to be made, the minimum cost to be paid in the interval $[k,N]$

---

* x_k is not white; however, this is irrelevant since it appears only in the cost function and not in state equation. The purpose of a whiteness assumption is to insure that the state is a Markov process, but in this case the state is deterministic and hence trivially Markov.

is clearly dependent on the state x'_k as well as $p(x_k/Z_k)$, the latter being referred to as the information state in Ref. 2. The resulting recursive search equation is identical to that of the combined optimization problem.

An alternative way to recognize this problem as a combined optimization problem consists of defining the composite state $x''_k = [x_k, x'_k]$ and of drawing the closed-loop block diagram of Fig. A-2, which is identical to that of the combined optimization problem.

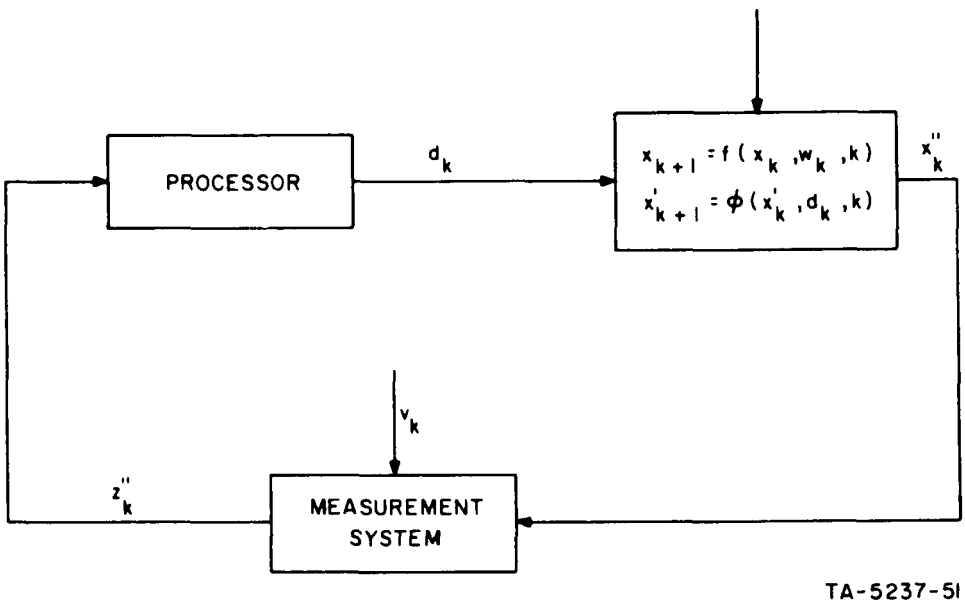

FIG. A-2 BLOCK DIAGRAM FOR CONSTRAINED DECISION PROCESS

x''_k = composite state of signal generating and decision process.

z''_k = composite measurement of x_k and x'_k.

D. Measures of Information

In all three cases, the minimization of the cost J requires knowledge of the conditional p.d.f. of state p(x/Z), and the optimum decisions d^* depend implicitly on this same p.d.f.

The function $p(x/Z)$ is consequently sufficient and necessary to summarize all the relevant information available. In special cases, such as linear Eqs. (1) and (2) and Gaussian p.d.f.'s $p(x_0)$, $p(w_k)$, $p(v_k)$, the function $p(x_k/Z_k)$ is Gaussian too and can be adequately and completely described by two sets of numbers, the mean $\hat{x}_{k/k}$ and the covariance $P_{k/k}$.

Thus, $p(x/Z)$ will be said to determine the quantity of information available, in the sense that it summarizes all available knowledge, both prior and collected.

The spread of the multivariate function $p(x_k/Z_k)$ can be approximately assessed by the convariance matrix or second moment,

$$P_{k/k} \triangleq \int_{x_k} (x_k - \hat{x}_{k/k})(x_k - \hat{x}_{k/k})^T \, p(x_k/Z_k) \, d\, x_k \tag{15}$$

or by the entropy, which is a single number

$$H_k = - \int_{x_k} p(x_k/Z_k) \log p(x_k/Z_k) \, d\, x_k \tag{16}$$

The rate at which the quantity of information increases with time is determined by comparing two successive p.d.f.'s such as $p(x_{k+1}/Z_{k+1})$ and $p(x_k/Z_k)$. This rate depends in a complex fashion on $p(x_k/Z_k)$ and on the random effects w_k and v_{k+1}, as can be illustrated by the linear Gaussian case (the Kalman-Bucy estimator) where the covariance $P_{k/k}$ is an exact measure of spread. In this case, the variances $P_{k+1/k+1}$ and $P_{k/k}$ can be shown to be related by the two recursive equations.

$$P_{k+1/k+1} = P_{k+1/k} - P_{k+1/k} H^T (H P_{k+1/k} H^T + R)^{-1} H P_{k+1/k} \tag{17}$$

$$P_{k+1/k} = \Phi P_{k/k} \Phi^T + \Gamma Q \Gamma^T \tag{18}$$

where Φ, Γ, and H are the matrices corresponding to Eqs. (1) and (2) and where Q and R are the covariances of w and v.

Although the function $p(x_k/Z_k)$ determines the quantity of information available, and is necessary to determine the cost J and the optimum decision d^*, it does not provide any clue concerning the sensitivity of J with respect to the quality of the data z_k [as measured for example by $p(v_k)$] or the amount of prior knowledge and thus does not help the designer in specifying the parameters of a measurement system or the accuracy of his model.

To determine this sensitivity, it is necessary to compute J in terms of the parameter defining the quality of the measurement system and the accuracy of the model. It is consequently the cost J which measures the value of a particular measurement system or a particular model. The dependence of J on each individual parameter gives a direct indication as to the desirability of changing this parameter and the dollar-cost entailed by such a change.

REFERENCES

1. E. Parzen, Modern Probability Theory and Its Applications, Wiley, 1960, p. 384.

2. L. Meier, "Combined Optimum Control and Estimation Theory," Contractor Report by SRI for NASA Ames Research Center on Contract NAS-2-2457, October 1965.

3. R. C. K. Lee, Optimal Estimation, Identification and Control, Chapter 3, MIT Press, 1964.

APPENDIX B

OPTIMUM QUANTIZATION

A. INTRODUCTION

In Sec. B of Ref. 6 a procedure is presented for optimally designing a quantizer with a fixed number of output levels. Both the values of the output levels and the range of the input to which each level corresponds are determined. The criterion for the design is that some measure of the average difference between the quantizer input and the quantizer output is minimized.

In this procedure the quantizer is viewed as a system element which introduces errors because the output can take on only finite numbers of values while the input can vary over a continuous range. The purpose of the design procedure is to minimize these errors.

This viewpoint is very appropriate for considering the information requirements of certain classes of systems. For example, if the system under consideration is an open-loop measurement device, the errors which the quantizer introduces are in a broad sense, related to the information lost in passing through the device. By designing the quantizer so as to minimize these errors, the information transfer through the device is maximized.

When the quantizer is introduced as an element of a closed-loop dynamic system, a number of additional questions arise. Many such questions were stimulated in connection with the example worked out in Sec. C of Ref. 6. The purpose of this present appendix is to answer these questions and to show clearly how optimum quantization applies in closed-loop systems.

One fact which was not pointed out in Sec. C of Ref. 6 is that the quantizer designed there actually performs three functions; namely, estimation, control and quantization.

In this present appendix, it is shown that under certain conditions the optimum combined system is found by optimizing the three

functions separately. This result is analogous to the separation of optimal control and optimal estimation found by Kalman,[1] Gunckel,[2] et al. for the linear, Gaussian, squared error case. The importance of this result is that in systems which have the appropriate properties, the optimum design procedure can be used directly without affecting the optimality of the total system. Even in cases where this result is not obtained, the procedure of separately designing the units generally yields a good approximation to the optimal system.

The remainder of the appendix consists of four sections. The first section presents a review of the design procedure. The second section formulates the problem of optimum quantization in a closed-loop dynamic system and discusses the difficulties that arise. The third section contains a proof of the separability of estimation, control, and quantization, for certain systems. Finally, the fourth section summarizes the work and draws some conclusions about its applicability.

B. A COMPUTATIONAL PROCEDURE FOR OPTIMUM QUANTIZER DESIGN

The formulation of the problem is based on the quantized characteristic shown in Fig. B-1. The quantizer input, x, is allowed to take on a continuous range of values. The probability density function of x, p(x), is assumed to be known. The quantizer output, q(x), is allowed to take on only a finite number of values, N + 1. The parameters which can be adjusted for optimum design are the (N + 1) discrete values of the quantizer output, $c_0, c_1, \ldots, c_N$, and the N points at which the output changes from c_{i-1} to c_i, denoted as $d_1, d_2, \ldots, d_N$.

The optimum quantizer design problem for the static case was formulated in Ref. 6 as a generalization of Tou's work.[3] Since the publication of Ref. 6, a Ph.D. thesis at Stanford[4] has appeared in which a similar problem is formulated and similar design equations are obtained. In Ref. 5, these latter equations are also derived, and a computational procedure analogous to the one described in this section is given.

The most general performance criterion that can be considered in a static situation is the expected value of some function of x and the

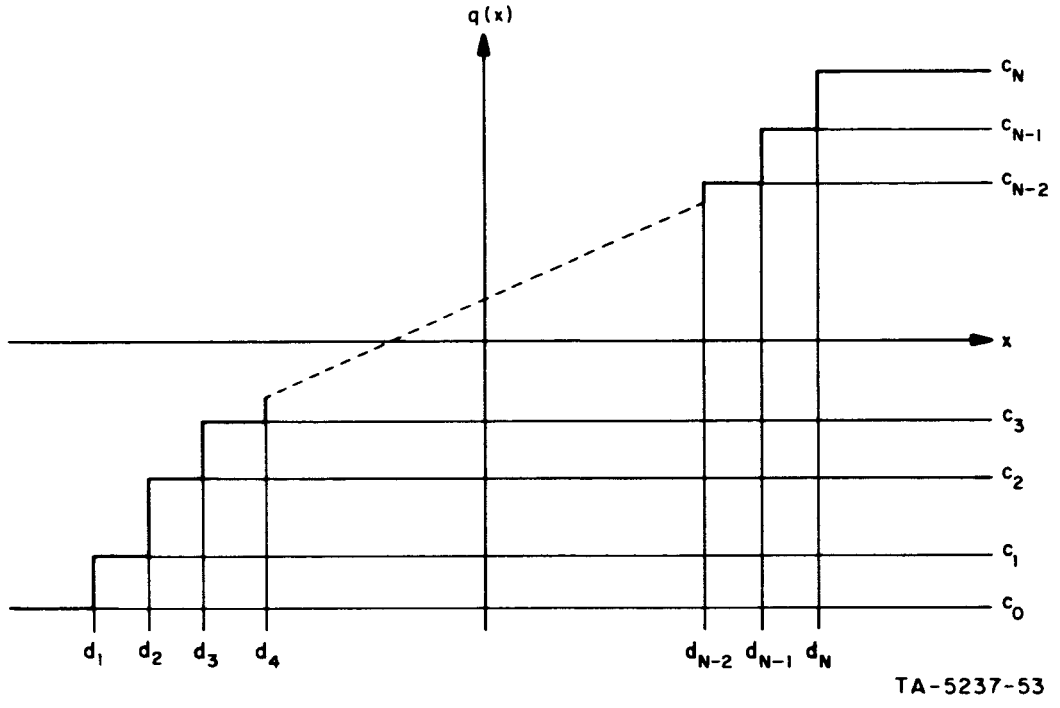

FIG. B-1 QUANTIZER CHARACTERISTIC

corresponding quantizer output. This function can be written as

$$J = \int_{-\infty}^{d_1} g(x,c_0)\ p(x)\ d\,x + \sum_{i=1}^{N-1} \int_{d_i}^{d_{i+1}} g(x,\ c_i)\ p(x)\ d\,x + \int_{d_N}^{\infty} g(x,\ c_N)\ p(x)\ d\,x \qquad (1)$$

In most cases the objective is to find a quantizer design that minimizes the expected value of a function of the error between the input and the corresponding output. In these cases the performance criterion becomes

$$J = \int_{\infty}^{d_1} g(x - c_0)\ p\ (x)\ d\ x$$

$$+ \sum_{i=1}^{N-1} \int_{d_i}^{d_{i+1}} g(x - c_i)\ p\ (x)\ d\ x + \int_{d_N}^{\infty} g(x - c_N)\ p\ (x)\ d\ x \qquad (2)$$

If d_0 is defined to be $-\infty$ and d_{N+1} to be $+\infty$, then Eq. (2) can be rewritten

$$J = \sum_{i=0}^{N} \int_{d_i}^{d_{i+1}} g(x - c_i)\ p(x)\ d\ x \qquad (3)$$

It is generally assumed [Refs. 4 and 5] that

$$\begin{aligned} g(o) &= 0 \\ g(\epsilon) &\geq 0, \text{ all } \epsilon \\ g(\epsilon) &= g(-\epsilon) \end{aligned} \qquad (4)$$

Two examples of this type of performance criterion are the expected value of absolute error, in which

$$g\ (x-c_i) = |\ x - c_i\ | \qquad (5)$$

and the expected value of squared error, where

$$g\ (x-c_i) = (x-c_i)^2 \qquad (6)$$

The computational procedure developed in Memorandum 2 can be extended to the performance criterion in Eq. (1); however, for purposes of this discussion, it will be assumed that the criterion has the form of Eq. (3).

The design equations can be derived directly by taking partial derivatives of J in Eq. (3). Using the normal rules for differentiating with respect to limits of integration,

$$\frac{\partial J}{\partial d_i} = 0 = g\,(d_i - c_{i-1})\,p\,(d_i) - g(d_i - c_i)\,p(d_i)$$

$$i = 1,2, \ldots .N \tag{7}$$

Differentiating under the integral sign,

$$\frac{\partial J}{\partial c_i} = 0 = -\int_{d_i}^{d_{i+1}} \frac{\partial g(x - c_i)}{\partial c_i}\, p(x)\, d\,x$$

$$i = 0,1,2,\ldots,N \tag{8}$$

Eq. (7) can be rewritten as

$$g\,(d_i - c_{i-1}) = g\,(d_i - c_i) \tag{9}$$

If in addition to Eq. (4) it is assumed, (and this is a most realistic assumption) that g is a monotonically increasing function of its argument, then Eq. (9) has as its unique solution

$$|d_i - c_{i-1}| = |d_i - c_i| \tag{10}$$

Since $c_{i-1} \neq c_i$, d_i is determined as

$$d_i = \tfrac{1}{2}\,(c_{i-1} + c_i) \tag{11}$$

that is, d_i is exactly half-way between c_{i-1} and c_i.

Equation (8) can be rewritten as

$$\int_{d_i}^{d_{i+1}} g'(x - c_i)\, p(x)\, dx = 0 \qquad i = 0,1,2,\ldots N \tag{12}$$

where g' is the derivative of g with respect to its argument.

If d_i is determined by Eq. (11), then the (2N + 1) free design parameters have been reduced to (N + 1), namely, the values of the c_i. If d_i and d_{i+1} in Eq. (12) are replaced by the appropriate forms of Eq. (11), then the conditions that must be satisfied become

$$\int_{-\infty}^{\frac{1}{2}(c_0 + c_1)} g'(x - c_0)\, p(x)\, dx = 0 \tag{13a}$$

$$\int_{\frac{1}{2}(c_{i-1} + c_i)}^{\frac{1}{2}(c_i + c_{i+1})} g'(x - c_i)\, p(x)\, dx = 0 \tag{13b}$$

$$i = 1,2,\ldots,N-1$$

$$\int_{\frac{1}{2}(c_{N-1} + c_N)}^{\infty} g'(x - c_N)\, p(x)\, dx = 0 \tag{13c}$$

It can be observed that Eq. (13a) depends only on c_0 and c_1. Consequently, if a guess is made of c_0, then c_1 can be determined. Proceeding to the first of Eq. (13b) it is seen that it depends only on c_0, c_1, and c_2. Consequently, c_2 can be determined from c_0 and c_1. In general, the ith Eq. (13b) yields c_{i+1} in terms of the previously calculated values of c_i and c_{i-1}. The (N-1)th of these equations determines c_N. If Eq. (13c) is zero when the computed values of c_{N-1} and c_N are substituted into it, then the values c_0, c_1, ..., c_N are the optimal set. If not, then a different value of c_0 must be tried and the procedure repeated. The value of the integral in Eq. (13c) indicates what new value to try for c_0.

For the absolute error criterion in Eq. (5),

$$g' = \begin{cases} +1 & , \quad c_i \le x \le d_{i+1} \\ -1 & , \quad d_i \le x \le c_i \end{cases} \tag{14}$$

The design equations become

$$\int_{-\infty}^{c_0} p(x)\, dx = \int_{c_0}^{\frac{1}{2}(c_0 + c_1)} p(x)\, dx$$

$$\int_{\frac{1}{2}(c_{i-1} + c_i)}^{c_i} p(x)\, dx = \int_{c_i}^{\frac{1}{2}(c_i + c_{i+1})} p(x)\, dx$$

$$\int_{\frac{1}{2}(c_{N-1} + c_N)}^{c_N} p(x)\, dx = \int_{c_N}^{\infty} p(x)\, dx \tag{15}$$

For the squared error criterion, Eq. (6),

$$g' = 2(x - c_i) \tag{16}$$

and the design equations are

$$c_0 \int_{-\infty}^{\frac{1}{2}(c_0 + c_1)} p(x)\, dx = \int_{-\infty}^{\frac{1}{2}(c_0 + c_1)} x\, p(x)\, dx$$

$$c_i \int_{\frac{1}{2}(c_{i-1} + c_i)}^{\frac{1}{2}(c_i + c_{i+1})} p(x)\, dx = \int_{\frac{1}{2}(c_{i-1} + c_i)}^{\frac{1}{2}(c_i + c_{i+1})} x\, p(x)\, dx \qquad i = 1,2,\ldots(N-1)$$

$$c_N \int_{\frac{1}{2}(c_{N-1} + c_N)}^{\infty} p(x)\, dx = \int_{\frac{1}{2}(c_{N-1} + c_N)}^{\infty} x\, p(x)\, dx \tag{17}$$

In many problems the quantizer output is restricted to have maximum and minimum values, such as

$$a \leq q(x) \leq b \tag{18}$$

In this case, the values $c_0, c_1, \ldots, c_N$ as determined by the computational procedure may not satisfy Eq. (18). It is then necessary to modify the computational procedure to account for this constraint. In Ref. 6, it is shown that the optimal procedure in this case is to set $c_0 = a$ and $c_N = b$, and then to select $c_1, c_2, \ldots, c_{N-1}$ so that they satisfy

$$\int_{\frac{1}{2}(c_{i-1} + c_i)}^{\frac{1}{2}(c_i + c_{i+1})} g'(x - c_i)\, p(x)\, dx = 0 \qquad i = 1,2,\ldots(N-1) \tag{19}$$

Since c_0 is known, c_1 can be guessed and the computational procedure previously described can be applied. Note that in this case, the quantizer parameters that are determined, depend on the probability density function $p(x)$ only over the interval $a \leq x \leq b$.

The results of applying the computational procedure to some examples are shown in Tables I, II, and III. In the first two examples, the constraint $-1 \leq q(x) \leq 1$ is imposed. The total number of output levels is $(N + 1) = 7$ for both cases. The probability density functions are $p(x) = k' e^{-|x|}$ and $p(x) = k'' e^{-1/2x^2}$ respectively. The cost function is expected absolute error, $g(x - c_i) = |(x - c_i)|$ in both cases. The total expected absolute error is computed in each case. These values are compared with the value for uniform quantization. In the third example, the output is not constrained, the total number of output levels is $(N + 1) = 7$, the probability density is Gaussian with zero mean and unity variance, and the performance criterion is expected squared error. Again, the total error is computed and compared. Total error for the best uniform quantization, which can also be found by a similar computational procedure,[5] is also computed for comparison purposes.

Table I

$N + 1 = 7,\ c_o = -1,\ c_6 = 1$

$g(x - c_i) = |x - c_i|$

$p(x) = .81e^{-|x|},\ -1 \le x \le 1$

	Optimum Quantization	Uniform Quantization
c_0	-1.00	-1.00
c_1	-0.59	-0.67
c_2	-0.27	-0.33
c_3	0.00	0.00
c_4	0.27	0.33
c_5	0.59	0.67
c_6	1.00	1.00
Total Cost	0.0820	0.0856

Table II

$N + 1 = 7,\ c_o = -1,\ c_6 = 1$

$g(x - c_i) = |x - c_i|^2$

$p(x) = .585e^{-\frac{1}{2}x},\ -1 \le x \le 1$

	Optimum Quantization	Uniform Quantization
c_0	-1.00	-1.00
c_1	-0.62	-0.67
c_2	-0.30	-0.33
c_3	0.00	0.00
c_4	0.30	0.33
c_5	0.62	0.67
c_6	1.00	1.00
Total Cost	0.0810	0.0842

Table III

$$N + 1 = 7,\ c_1 = -\infty,\ c_6 = \infty$$

$$g(x - c_i) = (x - c_i)^2$$

$$p(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

	Optimum Quantization	Uniform Quantization
c_0	-2.03	-1.95
c_1	-1.19	-1.30
c_2	-0.56	-0.65
c_3	0.00	0.00
c_4	0.56	0.65
c_5	1.19	1.30
c_6	2.03	1.95
Total Cost	0.0440	0.0469

C. OPTIMUM QUANTIZATION IN DYNAMIC SYSTEMS

In Ref. 6 the computational procedure was extended to dynamic systems. It was found that the functions of observation, control, and quantization all are inter-related. The purpose of this section is to clarify how each of these functions is to be carried out and to discuss other complications which arise when quantizers are introduced into dynamic systems.

A problem which illustrates most of the difficulties that arise is the case of a feedback control system where the control is quantized. This situation is shown in Fig. B-2 where the control u to be quantized is a scalar. The plant is described by a system of nonlinear time-varying difference equations.

$$\underline{x}(k + 1) = \underline{g}[\underline{x}(k), u(k), \underline{w}(k), k] \tag{20}$$

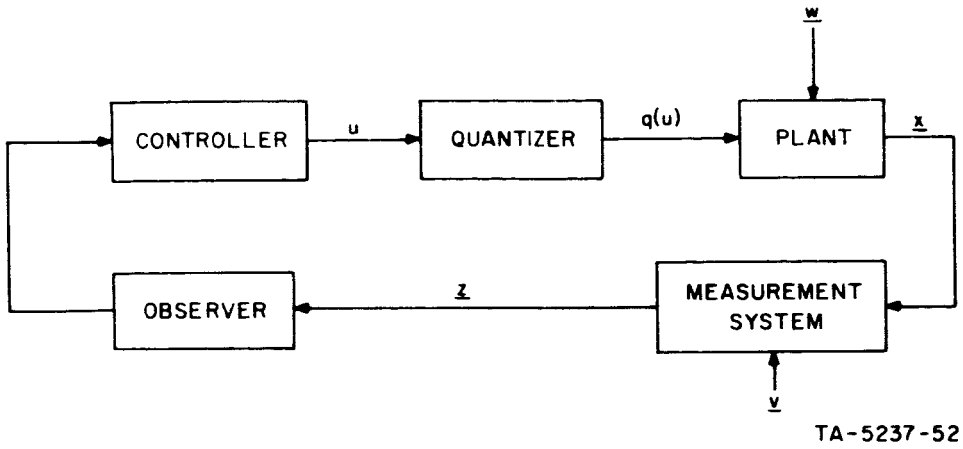

FIG. B-2 GENERAL DYNAMIC SYSTEM WITH QUANTIZED CONTROL

where

$\underline{x}$ (k) = n-dimensional state vector for system at time k

u (k) = scalar control at time k

$\underline{w}$ (k) = vector of random forcing functions on the system at time k

k = index of present time

The measurement system makes noise-corrupted observations of some or all of the states of the system according to the relation

$$\underline{z}\ (k) = \underline{h}\ [\ \underline{x}(k),\ \underline{v}\ (k),\ k\ ] \qquad (21)$$

where

$\underline{z}$ = m-dimensional vector of noisy measurements

$\underline{h}$ = m-dimensional vector functional where m is generally less than n

$\underline{v}$ (k) = vector of noisy signals that corrupt the measurements

The function of the block labelled "observer" is to process these measurements and extract as much information about the plant as possible. The function of the block labelled "controller" is to compute the "best" value of the input signal to the plant on the basis of the information that the observer provides. Best control is defined in general as

that which minimizes the expected value of some variational performance criterion.

$$J = E\left\{ \sum_{j=0}^{K} \ell\left[\underline{x}(j),\ u(j),\ \underline{w}(j),\ j\right] \right\} \tag{22}$$

where ℓ is a scalar functional.

The quantizer located between the output of the controller and the input to the plant makes control more difficult; instead of the plant being controlled by a signal selected from the continuum, the input is always taken from a finite set. The purpose of optimum quantization is to define this set so that the best performance that can be achieved under this restriction is actually obtained.

This problem is clearly more difficult than the static problem solved in Sec. B. The quantizer parameters must be selected to optimize not just some function of the instantaneous input and output of the quantizer, but instead a function of present and future values of the state of the plant, the controller output, the random forcing functions, and the measurement noise. Tou[3] shows that in some cases this problem can be solved by dynamic programming, but his procedure requires an $(N + 1)$-dimensional search for the $(N + 1)$ values of c_i at each stage. In Ref. 6 and related work, the $(N + 1)$-dimensional search is reduced to a one-dimensional search, and the class of problems that can be treated is extended. Nevertheless, the procedure is not always computationally practical. In the remainder of this appendix, more promising procedures based on combining the work outlined in Sec. B with the procedures for finding optimal control and optimal estimation will be pursued instead.

In order to carry out computations similar to those in Sec. B, it is necessary to have the probability density function of the input to the quantizer. If the structure of the quantizer, the plant, the measurement system, the observer, and the controller are all known,

and if the probability density functions for $\underline{x}$ (0), $\underline{v}(k)$, and $\underline{w}(k)$ are given, then this calculation is possible, at least in principle. However, a dilemma immediately arises: the quantizer is not fixed, yet its parameters must be known in order to carry out any calculations. This difficulty is circumvented by the following procedure: the probability density function of the quantizer input is computed assuming that the quantizer is replaced by a unity gain. The optimum quantizer is then designed. Using this new quantizer structure, the probability density function is re-computed. If it has changed significantly from the previous function, then a second quantizer design based on the new probability density function is computed. The procedure is repeated until the probability density function does not change significantly from one iteration to the next. Tou[3] has found that in general the probability density function based on approximating the quantizer by a unity gain is sufficent; in the cases where iterations are required, their number is small.

It is clear, however, that if this probability distribution is to be determined, then the controller and estimator must be fixed. If not, then the design problem becomes one of finding the optimum combined estimator-controller-quantizer combination. This problem is an extremely complicated one, which is even more difficult than the optimum combined estimator-controller problem treated by Meier in Vol. 1, Section III. Fortunately, as with Meier's work, there are a number of cases of practical importance in which a computationally feasible solution can be obtained. In addition, these simplifications lead to a useful procedure for treating the case of a fixed controller and fixed estimator as well. These problems are discussed in the next section.

D. COMBINED ESTIMATION, CONTROL, AND QUANTIZATION

The central result of this section is the following: if certain conditions are met, then the optimum combined observer-controller-quantizer design is obtained by first finding an optimum combined

controller-estimator and then synthesizing the optimum quantizer according to the procedure of Sec. B.

The conditions which must be satisfied in order to obtain this result are the following:

(1) The plant must be linear with zero-mean random forcing terms

$$\underline{x}\ (k+1) = \Phi \underline{x}(k) + \underline{d}\ u(k) + \Gamma\ \underline{w}\ (k) \tag{23}$$

where

Φ = n x n transition matrix

$\underline{d}$ = n-dimensional distribution vector

Γ = n x m distribution matrix for random forcing functions

$E[w(k)] = 0$

(2) The performance criterion must be a quadratic function of the present control and the next state.

$$J = E\ [\underline{x}^T(k+1)\ A\ \underline{x}(k+1) + b\ u^2\ (k)] \tag{24}$$

This criterion is a special case of the general variational criterion in Eq. (22). Tou[3] has shown that the optimum design obtained for this case is very close to the result for the summed quadratic criterion,

$$J = E\left[\sum_{j=k}^{\infty} [\underline{x}^T(j+1)\ A\underline{x}(j+1) + b\ u^2\ (j)\ ]\right] \tag{24a}$$

Furthermore, it is possible to extend the results of this section to the variational case, although the proof is somewhat involved.

(3) The optimum combined controller-observer is implemented by having the observer generate $\hat{\underline{x}}$ (k), the optimum estimate of the state vector, and then having the controller compute an optimum control signal treating this estimate as if it were in fact the true state. This is a well known result for the linear, Gaussian, squared error case; it can

be justified in a number of situations where these conditions do not hold exactly.

The proof that quantization can be separated from observation and control will be done by directly comparing the results of the separate optimization and the combined optimization.

1. Separate Optimization

In the case where the two optimizations are done separately, the optimal control is first computed under the assumption that the quantizer is a unity gain element. If the output of the observer is written as $\hat{\underline{x}}(k)$, then J can be re-written as

$$J = E\Big[[\Phi \hat{\underline{x}}(k) + \underline{d}u(k) + \Gamma \underline{w}(k)]^T A [\Phi\hat{\underline{x}}(k) + \underline{d}u(k) + \Gamma\underline{w}(k)] + bu^2(k) \Big]$$

$$= \Big[[\Phi\hat{\underline{x}}(k) + \underline{d}u(k)]^T A [\Phi\hat{\underline{x}}(k) + \underline{d}u(k)] + bu^2(k) + Q_1(k) \Big] \qquad (25)$$

where $Q_1(k) = E[\underline{w}^T(k)\ \Gamma^T A \Gamma\underline{w}(k)]$, a number which is independent of $\hat{\underline{x}}(k)$ and $u(k)$. The other terms involving $\underline{w}(k)$ vanish because $E[\underline{w}(k)] = 0$.

Expanding the terms in Eq. (25)

$$J = \hat{\underline{x}}^T(k)\ \Phi^T A \Phi\ \hat{\underline{x}}(k) + u^T(k)\ \underline{d}^T A \Phi\ \hat{\underline{x}}(k) + \hat{\underline{x}}^T(k)\ \Phi^T A\ \underline{d}\ u(k)$$

$$+ u^T(k)\ \underline{d}^T A\ \underline{d}\ u(k) + b\ u^2(k) + Q_1(k) \qquad (26)$$

Because $u(k)$ and $\underline{d}^T \Phi A\ \underline{x}(k)$ are both scalars, it follows that

$$J = \hat{\underline{x}}^T(k)\ \Phi^T A \Phi\ \hat{\underline{x}}(k) + u(k)\ [2\underline{d}^T A \Phi\ \hat{\underline{x}}(k)]$$
$$+ (b + \underline{d}^T A\underline{d})\ u^2(k) + Q_1(k) \qquad (27)$$

The minimization of J is accomplished by differentiating with respect to u and setting the result equal to 0,

$$\frac{\partial J}{\partial u} = 0 = 2\underline{d}^T A \Phi\ \hat{\underline{x}}(k) + 2(b + \underline{d}^T A\ \underline{d})\ u(k) \qquad (28)$$

The resulting optimal control, denoted by $\hat{u}(k)$, is

$$\hat{u}(k) = -(\underline{d}^T A\underline{d} + b)^{-1} (\underline{d}^T A\Phi)\hat{\underline{x}}(k) \tag{29}$$

which is the well-known solution for the case without quantization.

Now, consider the problem of designing a quantizer which minimizes the expected value of the square of the error between $\hat{u}(k)$ and $q[\hat{u}(k)]$. The probability density of $\hat{u}(k)$ is computed on the basis of the known controller and observer configurations and the known probability density functions for $\underline{x}(o)$, and $\underline{v}(k)$, and $\underline{w}(k)$. As in Ref. 3 the probability density function for $\hat{u}$ can be written as an average distribution over all possible sequences $\underline{x}(o)$, $\underline{w}(k)$, and $\underline{v}(k)$, $k = 0, 1, 2, \ldots, K$. The optimum quantization problem can be written as

$$J' = \underset{c_0, c_1, \ldots, c_N}{\text{Min}} \left\{ \underset{\hat{u}(k)}{E} \left[[\hat{u}(k) - q[\hat{u}(k)]^2 \right] \right\} \tag{30}$$

or, in terms of the state variables $\hat{\underline{x}}(k)$,

$$J' = \underset{c_0, c_1, \ldots, c_N}{\text{Min}} \left\{ \underset{\hat{\underline{x}}(k)}{E} \left[(d^T Ad + b)^{-1} (d^T A\Phi)\, \hat{\underline{x}}(k) + q \right]^2 \right\} \tag{31}$$

where the probability density of $\hat{\underline{x}}(k)$ is computed in the same way as $p[\hat{u}(k)]$. This problem, in the form of Eq. (30), can be solved by the computational procedures of Sec. B.

2. Combined Optimization

The formulation of the combined problem can now be given. Assumption (3) is used again to express the performance criterion as Eq. (25). The purpose of the quantizer design is to select levels so that the expected value of J in Eq. (25) is minimized. The problem becomes

$$J'' = \underset{c_0, c_1, \cdots c_N}{\text{Min}} \left\{ \underset{\hat{\underline{x}}(k)}{E} \left[(\Phi \hat{\underline{x}}(k) + \underline{d}q)^T A (\Phi \hat{\underline{x}}(k) + \underline{d}q) + bq^2 + Q_1(k) \right] \right\} \tag{32}$$

where the quantizer output q is substituted for u(k).

The equivalence of the two problems can be shown by demonstrating that the same set of values, $c_0, c_1, \ldots, c_N$ minimizes both of Eqs. (31) and (32). Expanding Eq. (32) and noting that q is a scalar,

$$J'' = \underset{c_0, c_1, \cdots c_N}{\text{Min}} \left\{ \underset{\hat{\underline{x}}(k)}{E} \left[\hat{\underline{x}}^T(k) \Phi^T A \Phi \hat{\underline{x}}(k) + 2q (\underline{d}^T A \Phi) \hat{\underline{x}}(k) + (b + \underline{d}^T A \underline{d}) q^2 + Q_1(k) \right] \right\} \tag{33}$$

$$= \underset{c_0, c_1, \cdots c_N}{\text{Min}} \left\{ \underset{\hat{\underline{x}}(k)}{E} \left[2q (\underline{d}^T A \Phi) \hat{\underline{x}}(k) + q^2 (\underline{d}^T A \underline{d} + b) + Q_2(k) \right] \right\}$$

where

$$Q_2(k) = Q_1(k) + \underset{\hat{\underline{x}}(k)}{E} \left[\hat{\underline{x}}(k) \Phi^T A \Phi \hat{\underline{x}}(k) \right]$$

which does not depend on q.

Expanding Eq. (31) and noting that $(\underline{d}^T A \underline{d} + b)$ is a scalar,

$$J' = \underset{c_0, c_1, \cdots c_N}{\text{Min}} \left\{ \underset{\hat{\underline{x}}(k)}{E} \left[(\underline{d}^T A \underline{d} + b)^{-2} (\underline{d}^T A \Phi \hat{\underline{x}}(k))^2 + 2(\underline{d}^T A \underline{d} + b)^{-1} \underline{d}^T A \Phi \hat{\underline{x}}(k) + q^2 \right] \right\}$$

$$= \underset{c_0, c_1, \cdots c_N}{\text{Min}} \left\{ \underset{\hat{\underline{x}}(k)}{E} \left[2q(\underline{d}^T A \underline{d} + b)^{-1} (\underline{d}^T A \Phi) \hat{\underline{x}}(k) + q^2 \right] + Q_3(k) \right\} \tag{34}$$

where

$$Q_3(k) = \underset{\hat{\underline{x}}(k)}{E} \left[(\underline{d}^T A d + b)^{-2} (\underline{d}^T A \Phi \hat{\underline{x}}(k))^2 \right],$$

which again does not depend on q.

Now, if $(\underline{d}^T A \underline{d} + b) = 0$, the optimal control fails to exist. Furthermore, $(\underline{d}^T A \underline{d}) > 0$ if A is positive definite, as is generally assumed. Also, b is always greater than or equal to 0. Therefore, Eq. (34) can be multiplied by $(\underline{d}^T A \underline{d} + b)$ without affecting the minimization.

$$J''' = \underset{c_0, c_1, \ldots c_N}{\text{Min}} \left\{ \underset{\hat{\underline{x}}(k)}{E} \left[2q (\underline{d}^T A \Phi) \hat{\underline{x}}(k) + q^2 (\underline{d}^T A \underline{d} + b) \right] + Q_4(k) \right\} \tag{35}$$

where

$$Q_4(k) = (\underline{d}^T A \underline{d} + b) Q_3(k)$$

Comparing Eqs. (33) and (35), we see that the expressions are identical except for the terms $Q_2(k)$ and $Q_4(k)$. However, since neither of these terms depends on q, the optimization is not affected in either case. Therefore, the solution for c_0, c_1, ..., c_N is the same for both Eqs. (33) and (35). This in turn implies that the solution to the two problems as expressed in Eqs. (31) and (32) is again the same. Therefore, the computational procedure can be applied to Eq. (30) to find the solution to the combined problem expressed in Eq. (32).

The consequence of this result in terms of computational requirements is that the optimum combined observer-controller-quantizer can be designed by first finding the best observer-controller design and then following it with an optimum quantizer designed by use of the computational procedure of Sec. B. The savings over the computational requirements for obtaining an optimal combined system directly are enormous.

As an illustration of the computational procedure, consider the two-dimensional example worked in Reference 6. The system equations are

$$\begin{bmatrix} x_1(t+1) \\ x_2(t+1) \end{bmatrix} = \begin{bmatrix} 1 & 2.995 \\ 0 & .96 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} 7 \cdot 10^{-8} \\ 2.1 \cdot 10^{-6} \end{bmatrix} u \tag{36}$$

where t has been normalized so that the unit time increment is 3 seconds [x (t + 1) is hence the state three seconds later than t]. The performance criterion is

$$J = \begin{bmatrix} x_1(t+1) & x_2(t+1) \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1(t+1) \\ x_2(t+1) \end{bmatrix} \tag{37}$$

therefore,

$$\Phi = \begin{bmatrix} 1 & 2.995 \\ 0 & 0.96 \end{bmatrix} \qquad \underline{d} = \begin{bmatrix} 7 \cdot 10^{-8} \\ 2.1 \cdot 10^{-6} \end{bmatrix}$$

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad b = 0 \tag{38}$$

In Section C of Ref. 6 the combined problem was solved directly. The computations were extremely involved and laborious. The equation finally reduced to

$$J'' = \min_{c_0, c_1, \ldots, c_N} \left\{ E_{\underline{\hat{x}}(k)} \left[1.065 \left[0.0314\, \hat{x}_1(k) + \hat{x}_2(k) \right] + 2.10 \cdot 10^{-6}\, q \right]^2 \right\} \tag{39}$$

Substituting directly into Eq. (35),

$$J' = \min_{c_0, c_1, \ldots c_N} \left\{ E_{x(k)} \left[2q \left[7 \cdot 10^{-8}\, \hat{x}_1(k) + 2.23 \cdot 10^{-6}\, \hat{x}_2(k) \right] + 4.41 \cdot 10^{-12}\, q^2 \right] + Q_4(k) \right\} \tag{40}$$

Expanding Eq. (39),

$$J'' = \min_{c_0,\, c_1, \ldots c_N} \left\{ E_{\underline{x}(k)} \left[2q \left(7 \cdot 10^{-8}\, \hat{x}_1(k) + 2.23 \cdot 10^{-6}\, \hat{x}_2(k) + 4.41 \cdot 10^{-12}\, q^2 \right] + Q_5(k) \right\} \tag{41}$$

where $Q_5(k)$ again does not depend on q. It is seen that Eqs. (40) and (41) differ only by the terms $Q_5(k)$ and $Q_4(k)$ which have no effect on the minimization. Therefore, the two problems have the same solution, and hence the combined problem can be solved by applying the computational procedure of Sec. B directly to Eq. (40), rather than using the laborious procedure of Section C of Ref. 6.

E. CONCLUSIONS

In Sec. B a computational procedure was developed for finding the optimum design of a quantizer in a static operating mode. The procedure reduces the problem of finding the (2N + 1) quantizer parameters to that of computing one single parameter. An iterative procedure that converges rapidly was developed for this one-dimensional search.

In Secs. C and D the case of a quantizer operating in a dynamic system was considered. The main result obtained is that in many cases, an overall optimal system can be designed by first synthesizing the observer-controller combination and then, using the procedure of Sec. B, determining the quantizer design. Even when this procedure does not yield an exact optimal, it is conjectured that the resulting design is quite close to optimal. The savings in computational requirements by doing the designs separately rather than simultaneously is considerable. Although attention was restricted to the case where the quantizer follows the controller, it is clear that the results can be extended to systems where quantizers appear elsewhere in the system.

The computational procedure of Sec. B can thus be applied to both static (open-loop) and dynamic (closed-loop) systems. The performance

criterion can be extended to include functions other than some measure of expected error. For example, the problem of designing a quantizer for which the variance of the output signal is closest to the variance of the input signal has been successfully formulated. Thus, the procedure as it stands is applicable to a large class of systems, and, by suitable modifications, it can be extended to include a great many other cases.

REFERENCES

1. Kalman, R. E., and R. S. Bucy, "New Results in Linear Filtering and Prediction Theory," Trans. A.S.M.E.J. Basic Eng., 1961.

2. Gunckel, T. L., "Optimum Design of Sampled-Data Systems with Random Parameters," Stanford Electronics Laboratory, T. R. 2102-2, Stanford, California, April, 1961.

3. Tou, J. T., Optimum Design of Digital Control Systems, Academic Press, New York, 1963.

4. Shaver, H. N. "Topics in Statistical Quantization," Ph. D. Thesis, Dept. of Electrical Engineering, Stanford University, Stanford, California, and SRI Report, April 1965.

5. Max. J., "Quantizing for Minimum Distortion," IRE Trans. PGIT, March, 1960.

6. Peschon, J., Larson, R. E., and Chen, A., "Optimum Design of Quantized Control Systems," SRI Memorandum 2 to Ames Research Center on Contract NAS-2-2457, February, 1965.

APPENDIX C

PRACTICAL COMPUTATION OF PROBABILITY DENSITIES IN DYNAMIC SYSTEMS

A. INTRODUCTION

In order to assess the degrading effects of random perturbations and measurement noise on the performance of a closed-loop system, it is usually necessary to compute the probability density function (p.d.f.) of the state x and sometimes the control u. Under transient conditions, these p.d.f.'s can be derived analytically, as is well known. In the general nonlinear non-Gaussian case recursive procedures, which are often much more efficient than Monte Carlo simulations, can be obtained by straightforward application of probability theory and numerical analysis.

It is the purpose of the present memorandum to show how these recursive equations are practically derived and how performance is computed from the p.d.f.'s thus obtained. No pretense is made to go into the details of the vast body of knowledge available in the field of stochastic differential and difference equations.

B. PROBLEM FORMULATION

A control system consisting of a plant, a set of sensors referred to as the measurement system, and a fixed controller is given (Fig. C-1). The plant is perturbed by white (uncorrelated in time) disturbances w of known p.d.f. and the sensors are affected by white noise v of known p.d.f. For mathematical convenience it is assumed that this system operates in discrete time, time being identified by the index k.

It is desired to calculate the performance J, which is a given function of x and u. Because of the perturbations v and w, the variables x and u entering into J are random. It is customary under those circumstances to calculate the expected value of the random variable J, $E\{J\}$, the expectation being over x and u.

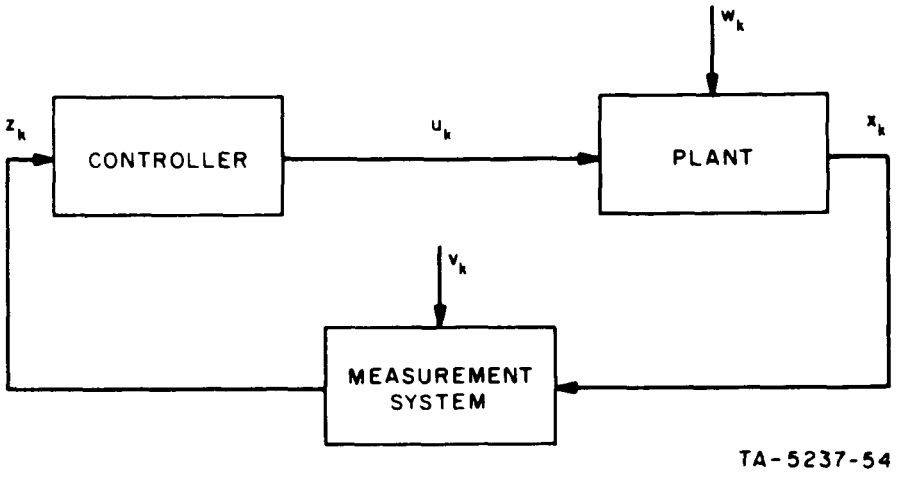

FIG. C-1 CONTROL SYSTEM WITH FIXED CONTROLLER

x_k = n dimensional state at time k

u_k = control

z_k = measurement system output

v_k = measurement system noise

w_k = disturbance

C. SYSTEM EQUATIONS

The plant, possibly expended to include sensor and actuator dynamics as well as shaping filters to convert ficticious white noise into actual colored noise, is described by

$$x_{k+1} = f(x_k, u_k, w_k, k) \tag{1}$$

The measurement system, whose internal dynamics have been removed as indicated above, is described by

$$z_k = h(x_k, v_k, k) \tag{2}$$

Alternatively, the measurement system may be given by the conditional p.d.f.

$$p\,(z_k \mid x_k,\ k) \tag{3}$$

The controller, which is fixed and not designed to utilize the received information best as is the case in combined optimization theory, is described by the algebraic equation

$$u_k = g\,(z_k,\ k) \tag{4}$$

where the time k accounts for possible time variations in the controller, as well as given and known command inputs.

By combining Eqs. (1), (2), and (4), the following difference equation is obtained

$$\begin{aligned} x_{k+1} &= f\left\{ x_k,\ g\left[h\,(x_k,\ v_k,\ k),\ k \right],\ w_k,\ k \right\} \\ &\triangleq F\left(x_k,\ w_k,\ v_k,\ k \right) \end{aligned} \tag{5}$$

Even though the variables u_k and z_k have been eliminated, the state Eq. (5) suffices to compute any transient motion resulting from an initial state x_o.

D. SYSTEM PERFORMANCE

Performance measures may be divided into static and dynamic measures. A static measure is defined at some time k as

$$\begin{aligned} J\left(x_k,\ h_k \right) &= \underset{x_k,\ u_k}{E}\left\{ \ell\,(x_k,\ u_k,\ k) \right\} \\ &= \int_{x_k} dx_k \int_{u_k} du_k\ p\,(x_k)\ p\,(u_k)\ \ell\,(x_k,\ u_k,\ k) \end{aligned} \tag{6}$$

A dynamic performance measure is defined over an interval [0,N], where N may be infinity, as

$$J\left(x_0, \ldots, x_N; u_0, \ldots, u_N\right) = E\left\{\sum_{k=0}^{N} \ell\ (x_k, u_k, k)\right\}$$

$$= \sum_{k=0}^{N} \left[\int_{x_k} dx_k \int_{u_k} du_k\ p\ (x_k)\ p\ (u_k)\ \ell\ (x_k, u_k, k)\right] \tag{7}$$

It is clear from Eqs. (6) and (7) that the calculation of performance requires knowledge of the p.d.f.'s $p(x_k)$ and $p(u_k)$.

E. NUMERICAL COMPUTATION OF $p(x_k)$ AND $p(u_k)$

In the general case, it is not possible to obtain closed form solutions for $p(x_k)$ and $p(u_k)$, but numerical computation by digital computer is always possible, though laborious for high-dimensional x, i.e., $n > 2$.

The first step consists of quantizing the variables x, v, w, and u. These quantized variables will be denoted by $\bar{x}$, $\bar{v}$, $\bar{w}$, and $\bar{u}$. With each quantized variable, there is associated a discrete probability distribution $p(\bar{x})$, $p(\bar{v})$, $p(\bar{w})$, and $p(\bar{u})$, respectively. For sufficiently small quantization increments, these discrete distributions approach the continuous p.d.f.'s p(x), p(v), p(w), and p(u).

1. Computation of $p(\bar{x}_k)$

From the chain rule

$$p\ (\bar{x}_{k+1}) = \sum_{\bar{x}_k} p\ (\bar{x}_{k+1} \mid \bar{x}_k)\ p\ (\bar{x}_k) \tag{8}$$

where the symbol $\sum_{\bar{x}_k}$ signifies summation over the quantized x_k space.

The distribution $p(\bar{x}_{k+1})$ is similarly obtained from

$$p\left(\bar{x}_{k+1}\right) = \sum_{\bar{x}_k} \sum_{\bar{v}_k} \sum_{\bar{w}_k} p\left(\bar{x}_{k+1} \mid \bar{x}_k, \bar{v}_k, \bar{w}_k\right) p\ (\bar{v}_k)\ p\ (\bar{w}_k)\ p\ (\bar{x}_k) \tag{9}$$

From the state Eq. (5), it is seen that x_{k+1} is no longer a random variable after x_k, v_k, w_k have been fixed, but assumes a well-defined value with probability one.

The computational procedure hence consists of selecting in sequence all possible combinations $\bar{x}_k$, $\bar{v}_k$, $\bar{w}_k$ with their associated probabilities, of computing the resulting $\bar{x}_{k+1}$ which occurs with probability $p(\bar{x}_k)\ p(\bar{v}_k)\ p(\bar{w}_k)$ and of accumulating these probabilities in the cell set aside for each of the $\bar{x}_{k+1}$. The distributions $p(\bar{v}_k)$ and $p(\bar{w}_k)$ are easily derived from the p.d.f's $p(v_k)$ and $p(w_k)$, whereas the distribution $p(\bar{x}_k)$ is known from the previous iteration.

2. Computation of $p(\bar{u}_k)$

By combining Eqs. (2) and (4), it follows that

$$u_k = g\left[\, h\,(x_k,\ v_k,\ k),\ k\,\right] \triangleq G\,(x_k,\ v_k,\ k) \tag{10}$$

Again, from the chain rule

$$p(\bar{u}_k) = \sum_{\bar{x}_k} \sum_{\bar{v}_k} p\left(\bar{u}_k \mid \bar{x}_k,\ \bar{v}_k\right)\ p\,(\bar{x}_k)\ p\,(\bar{v}_k) \tag{11}$$

where, for given $\bar{x}_k$ and $\bar{v}_k$, the resulting $\bar{u}_k$ is determined from Eq. (10) with probability one.

F. NUMERICAL COMPUTATION OF J

This calculation is carried out by straightforward application of Eq. (6) or (7) for the quantized variables $\bar{x}$ and $\bar{u}$. For example

$$J\left(x_k,\ u_k\right) \cong \sum_{\bar{x}_k} \sum_{\bar{u}_k} p\,(\bar{x}_k)\ p\,(\bar{u}_k)\ \ell\left(\bar{x}_k,\ \bar{u}_k,\ k\right) \tag{12}$$

G. STEADY-STATE SOLUTION

It frequently suffices to know the performance J after the system has settled down to a steady-state; that is, for $k \to \infty$ and in the absence of time-variations in Eqs. (1), (2), and (4), and in p(w) and p(v). To compute this performance, it is necessary to know the steady-state p.d.f. $p(x_k)$ for $k \to \infty$. This density function can often be computed directly as opposed to solving the recursive Eq. (9) over a sufficient number of increments.

The numerical procedure goes as follows:

Let $\bar{p}_k$ be a vector whose component $\bar{p}_k^i$ is the probability of the quantized state $\bar{x}_k$ being in the n-dimensional cube i at time k. Clearly, under steady-state conditions, the vectors $\bar{p}_k$ and $\bar{p}_{k+1}$ are identical.

But it is possible to relate $\bar{p}_{k+1}$ to $\bar{p}_k$ by means of a matrix S from the state Eq. (5). In effect

$$\bar{p}_{k+1}^{\,i} = \Pr\left(\bar{x}_{k+1} \text{ in } i\right) = \sum_j \Pr\left(\bar{x}_{k+1} \text{ in } i \mid \bar{x}_k \text{ in } j\right) \Pr\left(x_k \text{ in } j\right) \tag{13}$$

The number $\Pr(\bar{x}_{k+1}$ in i $\mid$ $\bar{x}_k$ in j) can be computed in a straightforward fashion from knowledge of the state equation and the distributions $p(\bar{w})$ and $p(\bar{v})$. This then determines the elements of S with the result that, in the steady state, $\bar{p}_k$ is given by

$$\bar{p}_k = S\,\bar{p}_k \tag{14}$$

The vector $\bar{p}_k$ which satisfies Eq. (11) is an eigenvector of S. In order for $\bar{p}_k$ to be a valid solution the following two conditions must obviously be met

$$\bar{p}_k^{\,i} \geq 0 \text{ for all } i \tag{15}$$

$$\sum_i \bar{p}_k^{\,i} = 1 \tag{16}$$

H. EXAMPLE

To illustrate this procedure, the following two-state example is considered.

Let the state x_k be in position 0 with probability $\bar{p}_k^0$ and in state 1 with probability $\bar{p}_k^1$. Let the transitions be represented by the flow diagram of Fig. C-2.

The elements of S are equal to the transition probabilities of Fig. C-2, e.g.

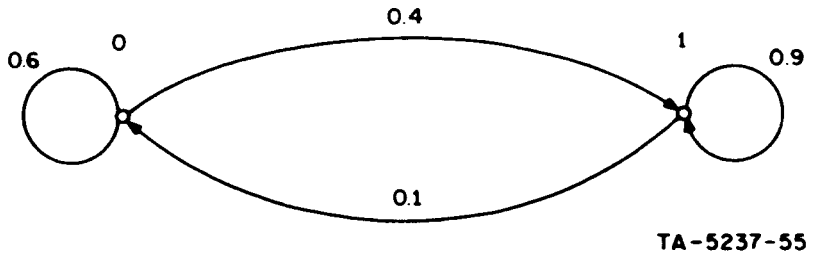

FIG. C-2 FLOW DIAGRAM OF EXAMPLE
(The numbers shown are the transition probabilities.)

$$S = \begin{bmatrix} 0.6 & 0.1 \\ 0.4 & 0.9 \end{bmatrix} \qquad (17)$$

The desired eigenvector $\bar{p}_k$ has components [1/5, 4/5] as can be easily verified by substitution into Eq. (17).

APPENDIX D

CLASSICAL PERFORMANCE MEASURES

A. USE OF CLASSICAL CONTROL THEORY

If the plants and all other components of a system are linear and constant then the state space descriptions given in Volume I may be replaced by a transfer function description by taking the Laplace transform of the differential equations. For such systems the classical methods of control theory such as block diagram manipulation, root locus, and sensitivity analysis can be fruitfully applied to the investigation of the effects of a non-ideal sensing system. Since such methods are well known and only their application in this context is novel, they will be illustrated by a simple example. A position servo was chosen for two reasons: Classical control theory is largely concerned with position control servos and the star-tracker example suggested by Ames is essentially a position servo.

1. Position Servo

In this discussion the simple position servo shown in Fig. D-1 will be considered. For the present K_2, K_3 are assumed to be zero and H(s) unity; later when the effect of using a rate sensor is considered these quantitites will take non-zero values.

2. Steady State Accuracy vs. Bias Errors

Inspection of the block diagram shows that the effect of a bias in the position sensor is a steady state error equal to the bias.

3. Location of Eigenvalues vs. Sensor Gain Changes

The effect on eigenvalue positions of changes from the nominal gain of the sensor may be investigated by root locus techniques[1] as shown in Fig. D-2.

A second means of analyzing the effect of gain changes is sensitivity analysis[2,3] as the following calculations indicate. The

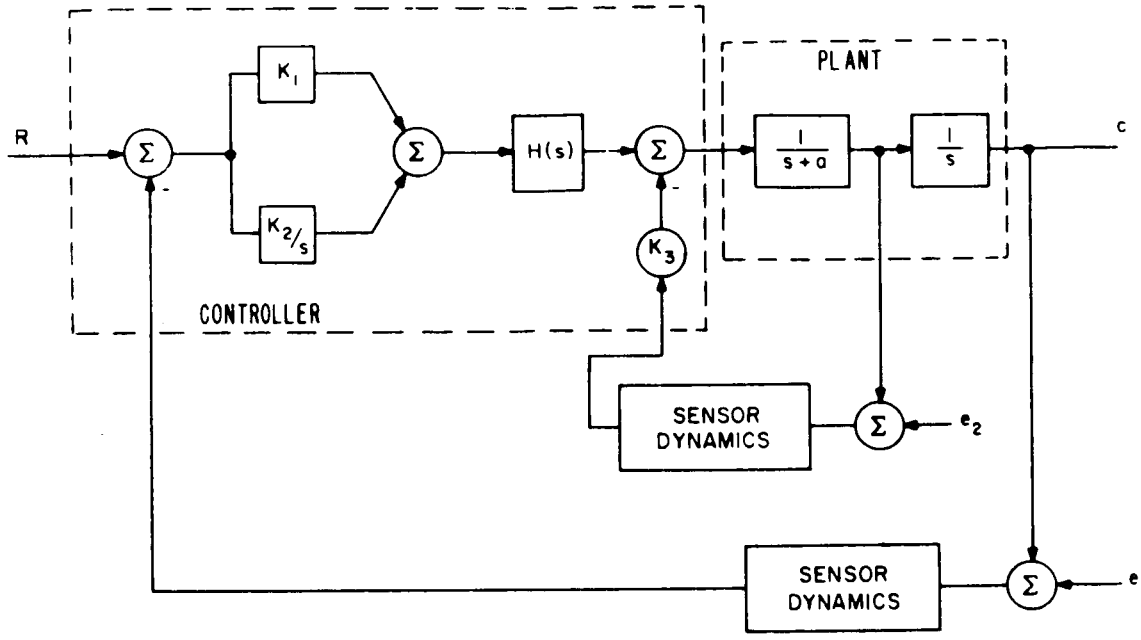

FIG. D-1 POSITION SERVO

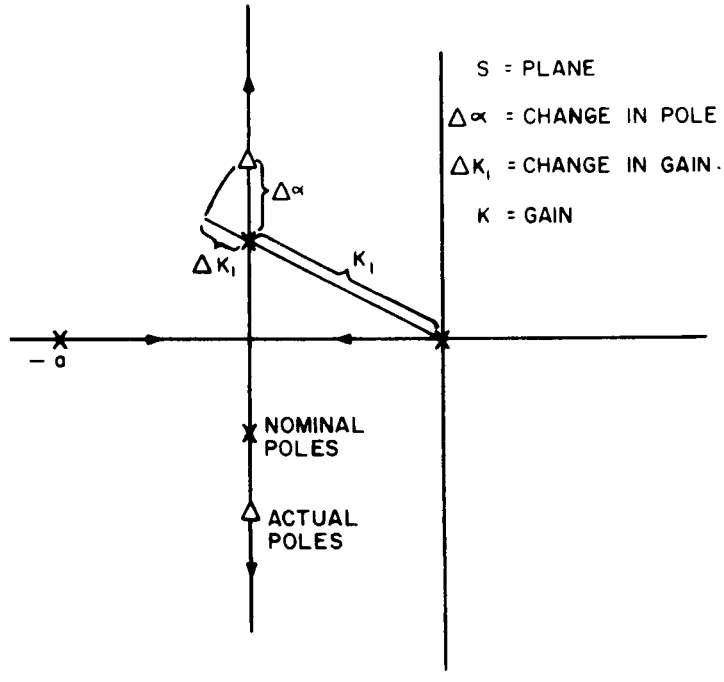

FIG. D-2 EFFECT OF GAIN CHANGES

characteristic equation of the system given in Fig. D-1 with $K_2 = K_3 = 0$ and $H(s) = 1$ is

$$s^2 + as + K_1 \triangleq (s + \alpha)(s + \beta) \tag{1}$$

If

$$K_1 \rightarrow K_1 + \Delta K \tag{2}$$

then:

$$\alpha \rightarrow \alpha + \Delta\alpha$$

$$\beta \rightarrow \beta + \Delta\beta$$

and

$$\Delta\alpha(s + \beta) + \Delta\beta(s + \alpha) + \Delta\alpha\Delta\beta = \Delta K \quad . \tag{3}$$

Since the coefficients s must be the same on both sides of (3)

$$\Delta\alpha = -\Delta\beta \tag{4}$$

and

$$-\Delta^2\alpha + (\beta - \alpha)\,\Delta\alpha = \Delta K \tag{5}$$

If $(\beta - \alpha)$ is not zero the second order terms may be neglected to give

$$\Delta\alpha = \frac{\Delta K}{\beta - \alpha} \quad \text{or} \quad \left|\frac{\Delta\alpha}{\alpha}\right| = \frac{\Delta K}{|\alpha||\beta - \alpha|} = \left|\frac{\Delta K}{K}\right| \left|\frac{\beta}{\beta - \alpha}\right| \tag{6}$$

For example if $\alpha = \frac{1}{2}(1 + \sqrt{3i})\sqrt{K}$ (i.e. 0.5 damping ratio)

$$\left|\frac{\beta}{\beta - \alpha}\right| = \frac{2}{\sqrt{3}} \tag{7}$$

In the singular case $\alpha = \beta$ (i.e. unity damping) (6) indicates that $\Delta\alpha$ is ∞; therefore the second order term in (5) cannot be ignored. Returning to (5) we get:

$$\frac{\Delta\alpha}{\alpha} = \sqrt{-\frac{\Delta K}{K}} \tag{8}$$

Given the value of expected gain changes, the poles changes can be calculated approximately using formulas such as (6); however, as the example shows, in singular cases misleading results may be obtained because second order terms have been neglected.

4. Sensor Dynamics

For the purpose of this discussion the sensor dynamics are taken to be a simple lag as shown in Fig. D-3; more complicated dynamics may be treated in the same manner.

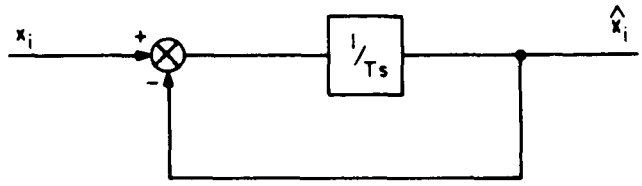

FIG. D-3 REPRESENTATION OF SENSOR DYNAMICS BY FIRST-ORDER LAG

The effect of sensor dynamics may be conveniently investigated by use of root locus as is illustrated in Fig. D-4.

The effect of changing T may also be investigated by the use of a root locus in which T varies and K_1 is fixed rather than K_1 varying and T fixed as in Fig. D-4. To do this the characteristic equation must be manipulated into suitable form. In our example the characteristic equation is

$$-1 = \frac{K_1}{s(s + a)(T s + 1)} \tag{9}$$

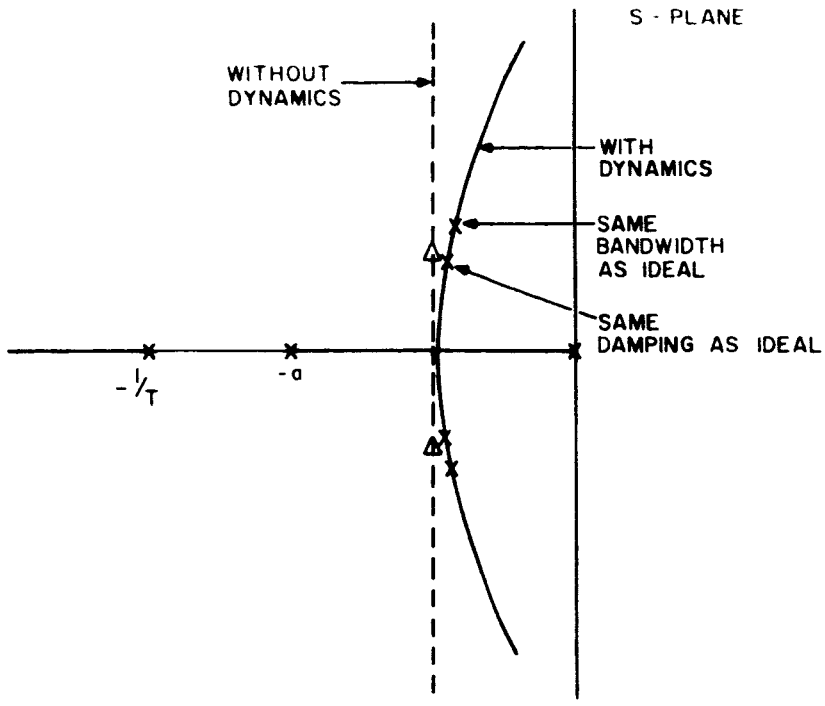

FIG. D-4 EFFECT OF SENSOR DYNAMICS

which is the form used for Fig. D-4. By manipulation we get

$$T s + 1 = -\frac{K_1}{s(s + a)}$$

$$-T s = \frac{s(s + a) + K_1}{s(s + a)}$$

$$-1 = \frac{T s^2 (s + a)}{K_1 + s(s + a)} \qquad (10)$$

The corresponding locus is shown in Fig. D-5.

An alternate method of analyzing the situation is sensitivity analysis: The transfer function $F_S(s)$ of the sensor given in Fig. D-2 is

$$G_S(s) = \frac{1}{T s + 1} \qquad (11)$$

For a perfect sensor T is zero; in a good sensor T is small compared to the systems dominant time-constant. Sensitivity analysis is applied to

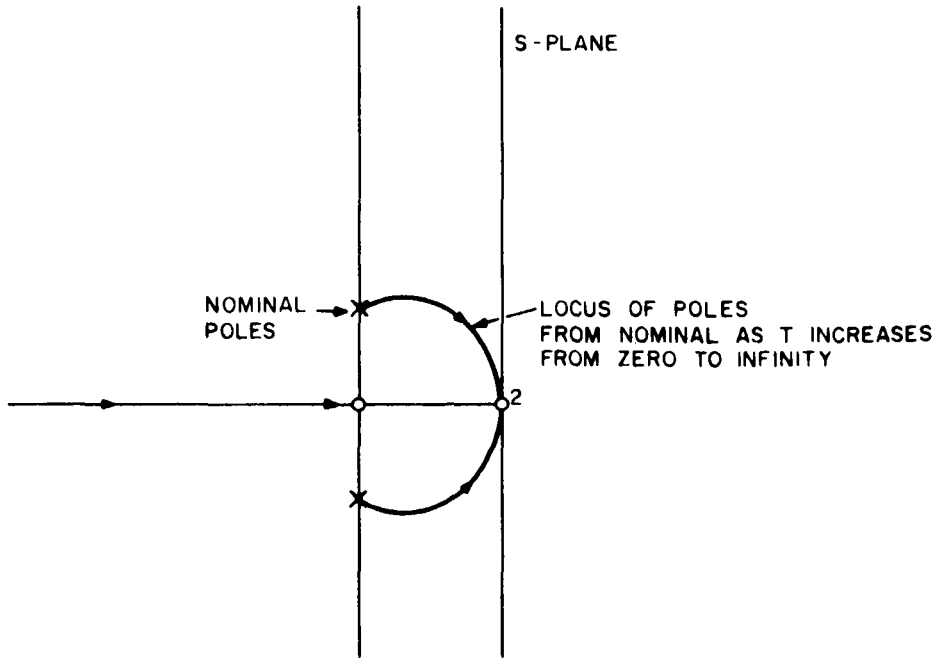

FIG. D-5 EFFECT OF VARYING SENSOR DYNAMICS

this problem by inserting $G_S(s)$ into the block diagram and calculating the sensitivity of the dominant poles to differences of T from its nominal value of zero.

6. Mean-Squared (M.S.) Error Due to Noise

The M.S. error added to the output due to noise may be calculated by finding the transfer function from noise to output, determining the power spectral density at the output using this transfer function and the power spectral density of the input, and finally integrating the expression for the noise in terms of the power spectral density by Cauchy's formula.[4] For example, assume that e_1, in Fig. D-1 is white noise with power spectral density

$$\Phi_N(\omega) = N \tag{12}$$

The transfer function T(s) from e_1, to c is

$$T(s) = \frac{K_1}{s^2 + as + K_1} = \frac{K_1}{(s+\alpha)(s+\beta)} \tag{13}$$

as can be seen by reference to Fig. D-6.

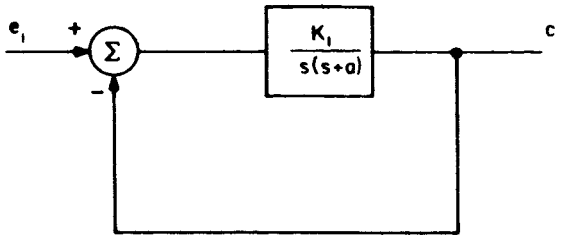

FIG. D-6 M. S. ERROR DUE TO NOISE

Reference to this figure and Fig. D-2 show that noise in the position servo enters the system in essentially the same manner as the reference input r. Using Eq. (13) we find the power spectral density $\Phi_c(\omega)$ of the output to be

$$\Phi_c(\omega) = |T(j\omega)|^2 \Phi_N(\omega) = \frac{K_1^2 N}{(\alpha^2+\omega^2)(\beta^2+\omega^2)}$$

$$= \frac{N K_1^2}{2(\alpha^2-\beta^2)} \left[\frac{-1/\alpha}{\alpha - j\omega} + \frac{-1/\alpha}{\alpha + j\omega} + \frac{1/\beta}{\beta - j\omega} + \frac{1/\beta}{\beta + j\omega} \right] \tag{14}$$

The mean square error $\overline{c^2}$ in the output is

$$\overline{c^2} = \frac{1}{2\pi j} \int_{-j\infty}^{j\infty} |T(j\omega)|^2 \Phi_c(\omega)\, d\omega$$

$$= \frac{K_1^2}{\alpha^2-\beta^2} \left[1/\beta - 1/\alpha \right] = \frac{K_1^2}{\alpha\beta(\alpha+\beta)}$$

$$= \frac{N K_1}{2a} \tag{15}$$

7. Use of Rate Sensor vs. Lead

Rate information, either measured by means of a sensor or calculated by means of a lead network, may be used to improve the transient response of the system. As can be seen from Fig. D-2, the position of the poles of the servo are constrained by "a"; however, the parameter "a" may be effectively increased by the use of rate information as may be seen from Fig. D-7 and Fig. D-8. Saturation effects always limit the values of $a + K_3$.

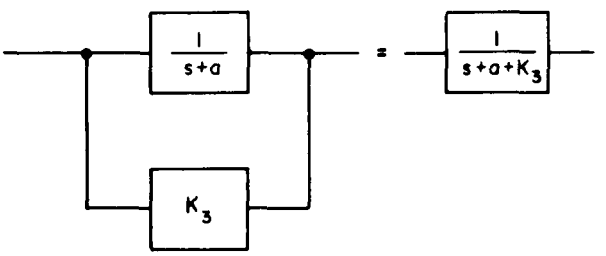

FIG. D-7 USE OF RATE FEEDBACK TO INCREASE a

Both methods of generating rate information have advantages and disadvantages: Use of a lead network is simple to implement whereas rate measurements are less affected by parameter variations and disturbance inputs (see Section D for an example). Either method for improving the transient response will increase steady state error due to noise input: the lead network by amplifying position sensor noise and rate feedback by introducing its own measurement noise.

Both low frequency noise and bias in a rate sensor may be effectively eliminated by use of integral control from the position sensor. If the gain K_2 in Fig. D-1 is zero then position bias and rate bias enter the system in a similar manner [see Fig. D-9(a)]; however, if K_2 is not zero then the integrator associated with it will take on a value such as to cancel the bias of the rate sensor [see Fig. D-9(b)]. In effect the position sensor is measuring the bias of the rate sensor.

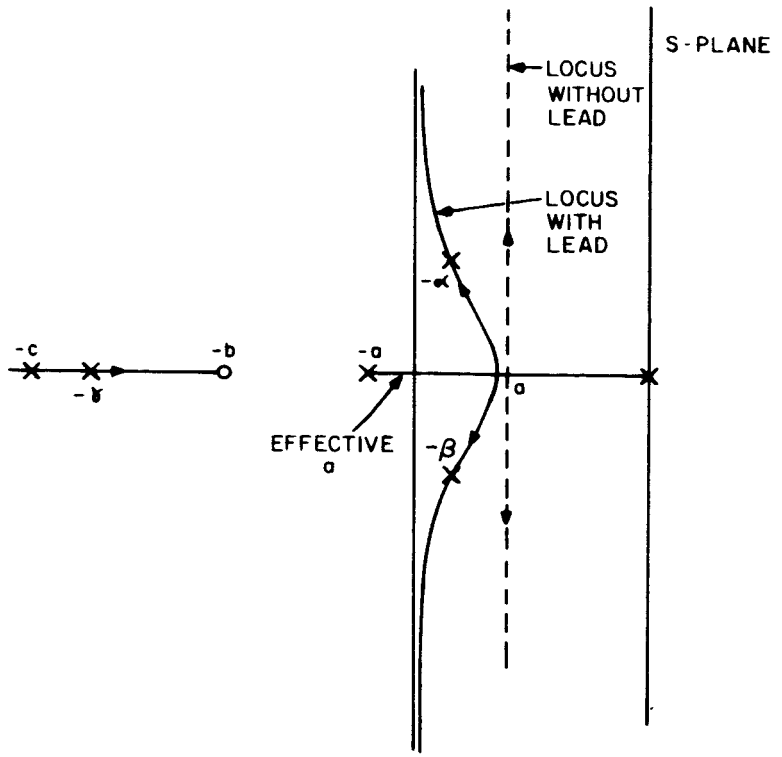

FIG. D-8 USE OF LEAD TO INCREASE A LEAD TRANSFER FUNCTION: $(s + b)/(s + c)$

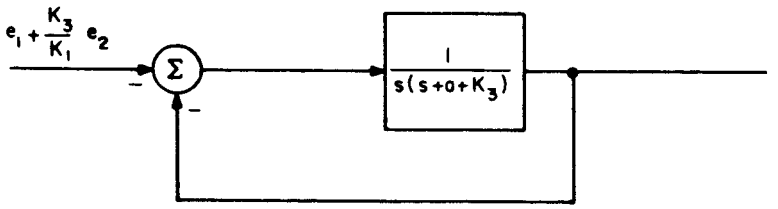

FIG. D-9(a) $K_2 = 0$

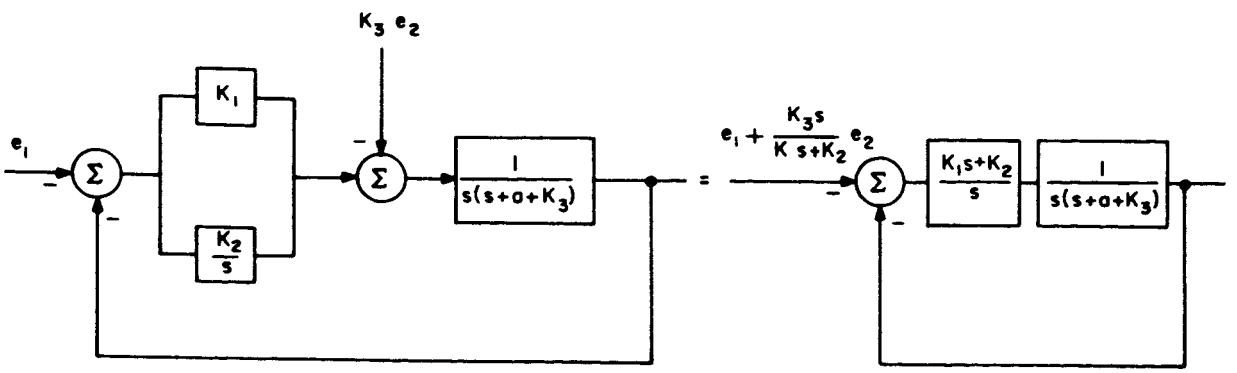

FIG. D-9(b) $K_2 \neq 0$

8. Sampling and Quantization

If the system contains a sampler it may be analyzed by the use of z-transform theory to calculate the response at sample instants and thus determine approximately the degradation in performance due to the sampler. If more accurate calculations are required the modified z-transform[5,6] can be used to calculate the response between sampling instants. Because the signal which is sent through a sampler must be limited to less than half the sampling frequency, the use of a sampler also implies a low pass input;[7] therefore the effect of a sampler is similar to that of lagging sensor dynamics, and the transient response will be adversely affected if the sampling rate is too slow.

Quantization is a much harder problem to treat using classical theory because the quantizer is a non-linear element. The effect of a quantizer is to add a time varying noise source[8] and non-linear bias to the system as Fig. D-10 indicates. The maximum bias error is 1/2 the quantization width for uniform quantizers.

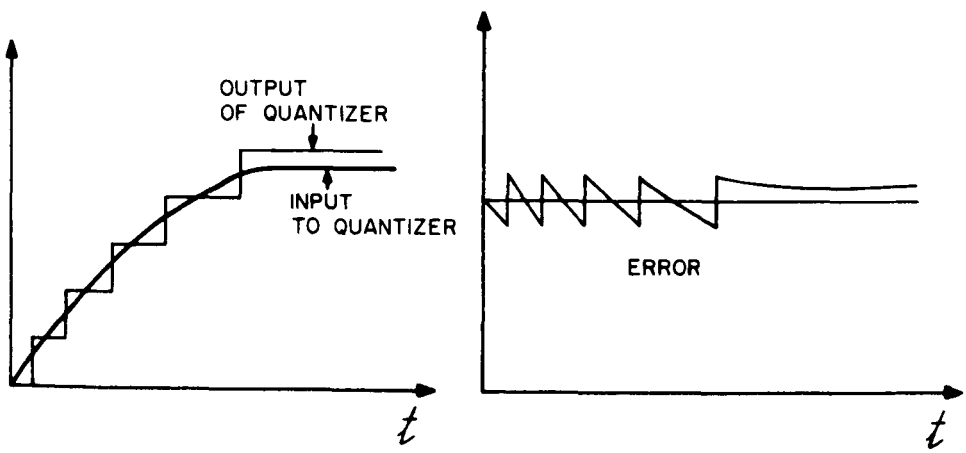

FIG. D-10 EFFECT OF QUANTIZER

B. STATE SPACE TECHNIQUES

1. Design and Analysis Procedure

By the use of the following two theorems, state space techniques may be used to synthesize controllers to meet classical system criteria. For the use of this method the state space representation of the information handling components is used.

THEOREM 1 (Kalman): Given a nth order system described by

$$\dot{x} = F x + g u \tag{16}$$

where the pair (F,g) is controllable.[9] Then a gain matrix k may be chosen so that $(F - g k^T)$ has specified eigenvalues i.e., if $u = k^T x$ then the system becomes

$$\dot{x} = (F - g k^T) x \tag{17}$$

and has arbitrary poles.

THEOREM 2 (Luenberger):[10] Given measurements

$$z = h^T x \tag{18}$$

of the system described by Eq. (17) and if the pair (F,h^T) is observable,[10] then an observer of order n-1 may be built to produce x arbitrarily rapidly i.e., the system

$$\dot{\hat{x}} = \hat{F} \hat{x} + \hat{g} u + q z$$

$$\hat{z} = \hat{H}_1 x + \hat{H}_2 z \tag{19}$$

may be chosen so that $\hat{z}$ approaches x with arbitrary speed. This latter property implies that $\hat{F}$ has arbitrary dynamics.

Based on these two theorems we can analyze the effect of information handling components as follows:

(a) A combined state is used for the plant and sensors

(b) k^T is chosen so that $F - k^T g$ has the proper eigenvalues

(c) The observer is chosen so that $\hat{F}$ has eigenvalues sufficiently removed from the dominant eigenvalues.

(d) The observer and k^T are combined to produce a controller by setting

$$u = k^T \hat{z} \tag{20}$$

The eigenvalues of the overall system will be those determined by the use of theorem 1 plus those determined by the use of theorem 2.

(e) The mean square error due to sensor noise is calculated using the power spectral densities of the noise as described in Sec. B-6.

The effect of the sensors may be determined by performing these calculations with and without the sensor defects present. Since the systems are designed so as to keep the poles constant, the effects of the sensor are measured by the increase in mean-square steady state error.

To illustrate this procedure, consider the following example:

Suppose that the plant is the one given in Fig. D-1, that only a position sensor is used, and that its only defect is measurement noise. For this system we have (for a = 1)

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= -x_2 + u \\ z &= x_1 \end{aligned}$$

$$F - g\,k^T = \begin{bmatrix} 0 & 1 \\ k_1 & -(k_2+1) \end{bmatrix} \tag{21}$$

The characteristic equation for this system is

$$\begin{aligned} \det\left[F - g\,k^T - s\,I\right] &= s(s + k_2 + 1) + k_1 \\ &= 0 = (s+\alpha)(s+\beta) \end{aligned} \tag{22}$$

If we specify the eigenvalues to be

$$-3 \pm 6j \tag{23}$$

then

$$k_1 = \alpha\beta = 9 + 36 = 45$$

and

$$k_2 = 5 \tag{24}$$

If the observer is given a pole at -15 it may be described by

$$\dot{\hat{x}} = -15\,\hat{x} + c_3 x_1 + u \tag{25}$$

$$\hat{z}_1 = x_1$$
$$\hat{z}_2 = c_1 x_1 + c_2 \hat{x} \tag{26}$$

where c_1, c_2 and c_3 are unknown parameters to be determined. Since $\hat{z}_1 - x_1$ is zero we only need to determine $\hat{z}_2 - x_2$. From (21), (25) and (26)

$$\dot{\hat{z}}_2 = c_1 \dot{x}_1 + c_2 \dot{\hat{x}}$$
$$= c_1 x_2 + -15c_2 \hat{x} + c_2 c_3 x_1 + c_2 u \tag{27}$$

and

$$\dot{\hat{z}}_2 - \dot{x}_2 = (1 + c_1) x_2 - 15c_2 \hat{x} + c_2 c_3 x_1 + (c_2 - 1) u \tag{28}$$

If we choose $c_2 = 1$, $c_1 = 14$ and $c_3 = -210$

$$\dot{\hat{z}}_2 - \dot{x}_2 = -15 (\hat{z}_2 - x_2) \tag{29}$$

and the error in estimating x_2 will die out as e^{-15t}.

When k^T and the observer are combined the following results are obtained:

From taking the Laplace transform of (25)

$$\hat{x} = \frac{u - 210\, x_1}{s + 15} \tag{30}$$

From (26)

$$\hat{z}_2 = 14\, x_1 + \hat{x} = \frac{14\, s}{s + 15}\, x_1 + \frac{1}{s + 15}\, u \tag{31}$$

From (24)

$$-u = 45\, \hat{z}_1 + 5\, \hat{z}_2$$

$$= 45\, x_1 + \frac{70\, s}{s + 15}\, x_1 + \frac{5}{s + 15}\, u \tag{32}$$

or when solved for u

$$-(s + 20)\, u = (115\, s + 675)\, x_1$$

$$u = -115\, \frac{s + 5.9}{s + 20}\, z \tag{33}$$

The closed loop transfer function is

$$T(s) = \frac{115\, s + 5.9}{(s + 20)(s + 1)s + 115\, s + 675}$$

$$= \frac{115\, (s + 5.9)}{s^3 + 21\, s^2 + 135\, s + 675}$$

$$= \frac{115\, (s + 5.9)}{(s^2 + 6\, s + 45)(s + 20)} \tag{34}$$

If the sensor noise is taken to be white with power spectral density

$$\Phi_N = 1 \tag{35}$$

then following the procedure given in Sec. B-6 the mean-square error $\bar{e}^2$ is

$$|e^2| = 9.7 \tag{36}$$

If instead of the lead network just designed the rate x_2 is measured, then with k_1 and k_2 as before the mean square error is

$$|e^2| = \left(1 + \frac{N}{9}\right) \frac{45}{6} = 7.5 + .83N \tag{37}$$

where the rate sensor noise is white with power spectral density

$$\Phi_N = N \tag{38}$$

By comparing (37) and (36), the trade-off between a lead network and a rate sensor may be made.

2. Sensitivity in State Space

If the complete system including the controller is written in state space notation then the sensitivity to parameter changes may be evaluated in an elegant, but not very well known, manner.[11] Suppose the overall state equation is

$$\dot{y} = A\,y \tag{39}$$

then the poles of the system are the eigenvalues λ_i of A. Let Y_i be the eigenvector corresponding to λ_i and V_i be the eigenvector of A^T corresponding to λ_i i.e.

$$A\,Y_i = \lambda_i\,Y_i$$

$$A^T\,V_i = \lambda_i\,V_i$$

or

$$V_i^T\,A = \lambda_i\,V_i^T \tag{40}$$

Consider a change dA of the matrix A; the corresponding changes $d\lambda_i$ and dY_i in λ_i and Y_i satisfy

$$d\,A\,Y_i + A\,dY_i = \lambda_i\,dY_i + d\lambda_i\,Y_i \tag{41}$$

where only first order terms are included. Multiplying (41) by V_i^T we get

$$V_i^T\,dA\,Y_i + V_i^T\,A\,d\,Y_i = \lambda_i V_i^T dY_i + d\lambda_i V_i^T\,Y_i \tag{42}$$

or, because of (40)

$$V_i^T\,dA\,Y_i = d\lambda_i\,V_i^T Y_i$$

or

$$d\lambda_i = \frac{V_i^T\,dA\,Y_i}{(V_i^T\,Y_i)} \tag{43}$$

If the norm $||dA||$ of dA is defined by

$$\|dA\| = \max_{y} \frac{y^T\, dA\, y}{(\bar{y}^T y)^{\frac{1}{2}}} \tag{44}$$

then

$$|d\lambda_i| \leq \frac{(\bar{V}_i^T V_i)^{\frac{1}{2}}\,(\bar{Y}_i^T Y_i)^{\frac{1}{2}}}{|V_i^T Y_i|}\,\|dA\| \tag{45}$$

note that the factor multiplying $||dA||$ is just the reciprocal of the cosine of the angle between V_i and Y_i.

As an example of the application of this procedure we will consider the situation analyzed in Sec. B-1 and assume that the sensor gain may change from its nominal value of 1. To take this effect into account, we modify (21) so that

$$z = k\, x_1 \tag{46}$$

A suitable state for the whole system is

$$y = \begin{bmatrix} x \\ \hat{x} \end{bmatrix} = \begin{matrix} x_1 \\ x_2 \\ \hat{x} \end{matrix} \tag{47}$$

For this system the state equation is

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= -115\, k\, x_1 - x_2 - 5\,\hat{x} \\ \dot{\hat{x}} &= -45\, k\, x_1 - 20\,\hat{x} \end{aligned} \tag{48}$$

Hence

$$A = \begin{bmatrix} 0 & 1 & 0 \\ -115k & -1 & -5 \\ -325k & 0 & -20 \end{bmatrix} \tag{49}$$

Since the nominal value of k is 1

$$dA = \begin{bmatrix} 0 & 0 & 0 \\ -115 & 0 & 0 \\ -325 & 0 & 0 \end{bmatrix} \Delta k \qquad (50)$$

If $\lambda_1 = -3 - 6j$ then

$$Y_1 = \begin{bmatrix} 1 \\ -3 - 6j \\ -17 - 6j \end{bmatrix} \qquad V_1 = \begin{bmatrix} -2 - 6j \\ 1 \\ -\frac{17 + 6j}{65} \end{bmatrix} \qquad (51)$$

and

$$V_1^T Y_1 = -1.1 - 8.8j$$

$$V_1^T \, d\,A\, Y_1 = (-30 + 30j)\,\Delta k \qquad (52)$$

therefore

$$|d\,\lambda_1| = \frac{|-30 - 30j|}{|1.1 + 8.8j|}\,|\Delta k| = 4.8|\Delta k| \qquad (53)$$

C. CONCLUSIONS

1. Transient vs. Steady-State

In general, the effects of imperfect information handling components such as sensors degrade either the transient response (for example sensor dynamics) or the steady-state response (for example sensor noise). Since by use of compensation and gain changes these two effects may be traded-off, the effect of imperfect components is to decrease steady-state performance for fixed transient performance and vice versa.

2. Use of Classical Performance Measures

"...the majority of pole-zero configurations encountered in the design of feedback control systems resolve to two or three poles and one or two finite zeroes...The possibility of such simplification means that the significant characteristics of the transient response can be rapidly estimated from the pole-zero configurations without recourse to the exact inverse transformation." The above quotation from Truxal[12] indicates that in a typical linear, constant, single-input, single-output control problem, the state of the system may be taken to be of order no higher than about 3. For such systems classical performance measures and analysis techniques are adequate.

3. Need for More Sophisticated Methods

(a) Nonlinearities and time variation. In general it is difficult to analyze nonlinear and time varying systems by means of classical control theory because this theory is largely based on Laplace transform theory.

(b) Multi-variable system. In systems with many inputs and outputs the number of state variables necessary for adequate description may be large. In such cases the trial and error processes of classical control theory are generally inadequate.

(c) For the preceding reasons as well as for its own sake, it would be desirable to have a general framework in which to analyze the effect of information handling components. Because such components may be described in state space notation such a framework is provided by otpimal control, estimation and identification theory. This method of viewing the problem is presented in the main body of the report.

D. SENSITIVITY OF RATE FEEDBACK VS. SENSITIVITY OF LEAD NETWORK

For the use of rate feedback as illustrated in Fig. D-10 the characteristic equation of the overall system given in Fig. D-4 is

$$s^2 + (a+K_3)\, s + K_1 \triangleq s^2 + (a+\beta)\, s + \alpha\beta = 0 \tag{54}$$

If $a \rightarrow a + \Delta a$ then $\alpha \rightarrow \alpha + \Delta\alpha$ and $\beta \rightarrow \beta + \Delta\beta$. When these changes are substituted into (54), higher order terms dropped, and coefficients of powers of s equated the following set of linear equations for $\Delta\alpha$ and $\Delta\beta$ result

$$\begin{aligned} \Delta\alpha + \Delta\beta &= \Delta a \\ \beta\Delta\alpha + \alpha\Delta\beta &= 0 \end{aligned} \tag{55}$$

with the solution

$$\begin{aligned} \Delta\beta &= -\frac{\beta}{\alpha}\Delta\alpha \\ \Delta\alpha &= \frac{\alpha}{\alpha - \beta}\Delta a \end{aligned} \tag{56}$$

For the situation shown in Fig. 8, the characteristic equation is

$$\begin{aligned} s^3 + (a+c)\, s^2 + (ac + K_1)\, s + K_1 b &= \\ (s+\alpha)(s+\beta)(s+\gamma) &= 0 \end{aligned} \tag{57}$$

where K_1 is chosen so that α and β are the same as above and where $|\gamma| >> |\alpha| = |\beta|$. Using the same process as above to find the linear equations for the perturbations $\Delta\alpha$, $\Delta\beta$ and $\Delta\gamma$ due to the perturbation Δa we get

$$\begin{aligned} \Delta\alpha + \Delta\beta + \Delta\gamma &= \Delta a \\ \Delta\alpha\,(\beta+\gamma) + \Delta\beta\,(\alpha+\gamma) + \Delta\gamma\,(\alpha+\beta) &= \Delta a\, c \\ \Delta\alpha\ \beta\gamma + \Delta\beta\ \alpha\ \gamma + \Delta\gamma\ \alpha\beta &= 0 \end{aligned} \tag{58}$$

with the solution

$$\Delta\gamma = \frac{-\gamma}{\alpha\beta}\,(\beta\,\Delta\alpha + \alpha\,\Delta\beta)$$

$$\Delta\beta = \frac{c\,\Delta a - \beta(1 - \frac{\gamma}{\alpha})\,\Delta\alpha}{\alpha(1 - \frac{\gamma}{\beta})}$$

$$\Delta\alpha = \frac{\alpha}{\alpha - \beta}\,(\frac{c - \alpha}{\gamma - \alpha})\,\Delta a \qquad (59)$$

Comparing (59) with (56) we see that for the same change in a, the change in α differs in the two cases by the factor $(c-\alpha)/(\gamma-\alpha)$. Because both c and γ are both real and $c > \gamma$ (see Fig. D-8) this factor always has magnitude greater than 1; therefore the rate feedback configuration is less sensitive to changes in the plant parameter "a" than the lead network configuration.

REFERENCES

1. J. G. Truxal, Control System Synthesis, McGraw-Hill, 1955, Chapter 4, pp. 221-277.

2. W. R. Perkins, "The Sensitivity of Feedback Control Systems to Parameter Variations," T. R. # 2107-1, Stanford Electronics Laboratory, Stanford, California.

3. Truxal, Op. Cit., Section 2.5, pp. 120-127.

4. Ibid, Section 8.1, pp. 454-460.

5. Ibid, Section 9.3, pp. 508-517.

6. Julius T. Tou, Digital and Sampled Data Control Systems, McGraw-Hill, 1959, Section 5.7, pp. 184-198.

7. Ibid, Chapter 3, pp. 69-92.

8. B. Widrow, "A Study of Rough Amplitude Quantization by Means of Nyquist Sampling Theory," Trans. IRE, Vo. CT-3, No. 4, Dec. 1956, pp. 266-276.

9. R. Kalman, "Controllability of Linear Dynamical Systems," Cont. to Differential Equations. Vol. I, No. 2, pp. 189-213.

10. D. G. Luenberger, "Observing the State of a Linear System," Trans. IEEE, Vol. MIL-8, No. 2, pp. 189-213.

11. Faddeev and Faddeeva, Computational Methods of Linear Algebra, Freeman, 1963 pp. 228-231.

12. J. G. Truxal, Ibid, p. 43.

13. J. Peschon, W. H. Foy, and L. Meier, "Information Requirements for Guidance and Control Systems," SRI Midterm Report to NASA Ames, Contract NAS 2-2457, March 1965.

APPENDIX E

INFORMATION-THEORETIC APPROACH

A. GENERALITIES

The collection of mathematical results known as Information Theory has been extremely fruitful in providing basic insight into communication processes. In the main, these results apply most directly to communication channels operating in a steady-state mode, in particular with constant probability distributions over the space of inputs. The most important conclusions apply to one-way channels in which arbitrarily large time delays are acceptable. The theory is based on the concept of entropy (i.e., the logarithmic function of the probability) as a measure of uncertainty. The great utility of this idea is that if two events are statistically independent the entropy of their joint occurrence is the sum of their individual entropies, in keeping with one's intuitive idea of the way in which uncertainties should behave.

When we turn to consideration of control systems, however, we find that many of these conditions simply do not apply. For example, the steady-state operation of a control system is often of trivial importance compared to its transient behavior, which will determine such a critical performance characteristic as response time. Under transient conditions the probability distribution of the input to a communication channel in the control path will certainly not be constant. Further, many control systems are quite sensitive to the value of a time delay in the signal path. Finally, many control systems at one point or another exhibit the operation of adding arithmetically two random variables to produce a third. It turns out that the entropy of the sum variable is not even calculable in general from the individual entropies of the two variables being added;* one needs the complete probability distributions of the

---

*This is easy to see. One need merely take one of the summands to be bimodal. For normally distributed variables the entropy of the sum is determined by the individual entropies, but this is a rather special case.

summand variables. In control system analysis, therefore, one often does not enjoy the analytical property that makes entropy such a very useful idea in the analysis of communication links.

Such considerations as these have rather deep implications when one comes to investigate the appropriate routes to be taken in the analysis of information requirements in guidance and control problems. A specific conclusion is that one probably must study, at the start, anyway, the complete probability distribution of a system quantity rather than one or a few of its real-valued properties.

B. PROBABILISTIC CONTROL PROBLEM

With this idea in mind, then, we set up a model system that is general enough to contain a variety of interesting control problems involving information handling elements and yet that possesses some useful analytical properties. Let us consider, specifically, a system with a state-vector $\bar{x}$ of n real elements with its evolution over the time period $0 \leq t < \infty$ described by the differential equations

$$\dot{\bar{x}} \overset{\text{def}}{=} \frac{d\bar{x}}{dt} = F\bar{x} + G\bar{\omega} + H\bar{y} \tag{1}$$

where F, G, H are matrices of (possibly) time-varying real elements. Take $\bar{\omega}$ to be a real vector of random elements, disturbances or additive noise, with the ensemble moments

$$\begin{aligned} <\bar{\omega}(t)> &= 0 \\ <\bar{\omega}(t_1)\omega^{-T}(t_2)> &= Q(t_1)\delta(t_1 - t_2) \end{aligned} \tag{2}$$

so that $\bar{\omega}(t)$ is zero-mean and white with instantaneous covariance matrix Q. Take $\bar{y}$ to be a feedback vector of m elements with dependence on the state $\bar{x}$ described by a conditional probability density $p_f(\bar{y}|\bar{x})$ so that

$$p_f(\bar{u}|\bar{v})d\bar{u} = \text{Prob.}\ (u_1 < y_1 < u_1 + du_1, \ldots, u_m < y_m < u_m + du_m$$

$$\text{given that state is}\ \bar{v})$$

and

$$\int\!\!\ldots\!\int d\bar{y}\ p_f(\bar{y}|\bar{x}) = 1$$

and define the ensemble conditional moments

$$\bar{m} = \bar{m}(\bar{x}) \overset{\text{def}}{=} \int\!\!\ldots\!\int d\bar{y}\ \bar{y}\ p_f(\bar{y}|\bar{x})$$

$$S = [S_{jk}(\bar{x})] \overset{\text{def}}{=} \int\!\!\ldots\!\int d\bar{y}\ \{\bar{y} - \bar{m}\}\{\bar{y} - \bar{m}\}^T\ pf(\bar{y}|\bar{x}) \qquad (3)$$

We suppose that the $\bar{y}$ vectors selected from the feedback distribution at different times are statistically independent, i.e.,

$$\langle\{\bar{y}(t_1) - \bar{m}_1\}\{\bar{y}(t_2) - \bar{m}_2\}|\bar{x}(t_1),\ \bar{x}(t_2)\rangle = 0$$

with similar expressions for all higher moments. Also, that $\bar{y}$ is statistically independent of $\bar{\omega}$. Thus, the feedback $\bar{y}$ is probabilistically dependent on the state $\bar{x}$, is "white" in time, and has instantaneous average $\bar{m}$ and covariance matrix S. Equation (1) is linear except for possible nonlinear dependence of the feedback $\bar{y}$ on the state $\bar{x}$; specifically, we do not assume that $\bar{m}(\bar{x})$ and $S(\bar{x})$ are linear in the elements of $\bar{x}$.

The quantity we shall consider to be of basic interest here is the probability density $p(\bar{x},t)$ of the system state in $\bar{x}$ space at any instant of time; i.e., we define

$$p(\bar{v},t)d\bar{v} = \text{Prob}(v_1 \leq x_1\ (t) \leq v_1 + dv_1, \ldots v_n \leq x_n(t)$$

$$\leq v_n + dv_n \text{ with } t \text{ given}$$

and

$$\int\cdots\int d\bar{x}\; p(\bar{x},t) = 1$$

The density $p(\bar{x},t)$ is a real function of $n+1$ real variables. In what follows we shall make the following assumptions about this function:

(a) Smoothness--The second partial derivatives of p with respect to the elements of $\bar{x}$ and its first partial derivatives with respect to t are all continuous.

(b) Boundary Behavior--When $|x_i| \to \infty$, $p(\bar{x},t)$ and its first partial derivatives with respect to the elements of $\bar{x}$ all vanish more rapidly than any finite power of x_j, for all i, j, and finite t.

The system of Eq. (1) can usually be set up so that these assumptions are physically plausible; no attempt has been made here at their mathematical justification. Finally, we assume for the present that the initial state probability density $p(\bar{x},0)$ is known. This completes our problem formulation.

The definition of a state vector and our "whiteness" specifications for $\bar{w}$ and $\bar{y}$ insure that the time evolution of the system state, $\bar{x}(t)$, will be a Markov process. Then, as we show in Section C of this Appendix, the state probability density $p(\bar{x},t)$ must obey a particular linear partial differential equation, the Fokker-Planck equation:

$$\frac{\partial p}{\partial t} = -\sum_{k=1}^{n} \frac{\partial}{\partial x_k}(\alpha_k p) + \frac{1}{2}\sum_{k=1}^{n}\sum_{j=1}^{n} \frac{\partial^2}{\partial x_k \partial x_j}(\beta_{kj} p) \qquad (4)$$

where we have defined the n-vector

$$\bar{\alpha} = \bar{\alpha}(\bar{x}) \stackrel{\text{def}}{=} F\bar{x} + H\bar{m}$$

and the matrix

$$[\beta_{kj}] = [\beta_{kj}(\bar{x})] \stackrel{\text{def}}{=} GQG^T + HSH^T$$

This equation is fundamental to all that follows here. It shows directly that the evolution of $p(\bar{x},t)$ can be described completely in terms of $\bar{m}$ and S, the first and second moments of the feedback term $\bar{y}$, and that we no longer need be concerned with the complete feedback conditional density $p_f(\bar{y}|\bar{x})$. This is an important simplification.

1. Entropy Variation

In theory, everything of present interest could be determined by integrating the Fokker-Planck equation; this, however, is a formidable task. A more useful approach is to write various quantities of interest in terms of integrals of the $p(\bar{x},t)$ density and look for an ordinary differential equation that will describe the quantity. For example, we can write the entropy of the state at any time as

$$\mathcal{H}(t) \stackrel{\text{def}}{=} -\int\cdots\int d\bar{x}\; p(\bar{x},t)\log_e(p(\bar{x},t))$$

Since this quantity can be interpreted as a measure of the uncertainty as to system location in $\bar{x}$ space it is of considerable interest. Differentiating with respect to time, we get

$$\dot{\mathcal{H}} = -\int\cdots\int dx\{1 + \log_e p(x,t)\}\frac{\partial p(\bar{x},t)}{\partial t}$$

Substitute for the partial derivative with respect to time the right-hand side of Eq. (4), the Fokker-Planck equation, and this is

$$\dot{\mathcal{H}} = \sum_{k=1}^{n}\int\cdots\int d\bar{x}\{1 + \log_e p\}\frac{\partial}{\partial x_k}(\alpha_k p)$$

$$-\frac{1}{2}\sum_{k=1}^{n}\sum_{j=1}^{n}\int\cdots\int d\bar{x}\{1 + \log_e p\}\frac{\partial^2}{\partial x_k \partial x_j}(\beta_{kj} p)$$

where the integrals now contain only partial derivatives with respect to the variables of integration. Apply integration by parts to the integrals in the first sum and invoke the boundary-behavior assumption (b) on $p(x,t)$, so that

$$\dot{\mathcal{H}} = \sum_{k=1}^{n} \int..\int d\bar{x}\; p(\bar{x},t)\; \frac{\partial}{\partial x_k}\, \alpha_k(\bar{x}) \qquad (5)$$

$$- \frac{1}{2} \sum_{k=1}^{n} \sum_{j=1}^{n} \int..\int dx\{1 + \log_e p\} \frac{\partial^2}{\partial x_k \partial x_j}\; (\beta_{kj} p)$$

We have found no way to simplify the second sum that gives useful results, and so this is as far as we can go without particularizing our system model. This is unfortunate, since the second sum contains the effects of noise, disturbances, and probabilistic feedback and is therefore the more interesting term. We can get an interesting special result, however, by taking the disturbances to be zero and the feedback to be linear and deterministic so that we have

$$\bar{\omega} = 0 \qquad Q = 0$$

$$\bar{m}(\bar{x}) = M\bar{x} \qquad S = 0$$

with M a matrix not dependent on $\bar{x}$, and thus

$$\bar{\alpha}(\bar{x}) = [F + HM]\bar{x} \quad [\beta] = 0$$

The only probabilistic effects then are those introduced by the initial probability density $p(\bar{x},0)$, so the system evolves deterministically from an ensemble of initial states. Equation (5) for the entropy becomes, in this case,

$$\dot{\mathcal{H}} = \text{trace}\,[F + HM] \qquad (6)$$

which is a simple and intriguing result. Consider an example in which the system is described by

$$\ddot{z} + 2\gamma\dot{z} + (w^2 + \gamma^2)z = 0 \qquad t \geq 0$$

and take for the state space the z and $\dot{z}$ elements; the system can be written

$$\dot{\bar{x}} = \frac{d}{dt}\begin{bmatrix} z \\ \dot{z} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -(w^2+\gamma^2) & -2\gamma \end{bmatrix}\begin{bmatrix} z \\ \dot{z} \end{bmatrix} = F\bar{x}$$

so that, by Eq. (6)

$$\mathcal{H}' = -2\gamma$$

$$\mathcal{H}(t) = \mathcal{H}(0) - 2\gamma t$$

If $\gamma > 0$ the system is damped, and we find that the entropy decreases; if $\gamma < 0$ the system diverges and the entropy increases. The case $\gamma = 0$ corresponds to a physical system in which energy is conserved and Liouville's theorem in statistical mechanics implies that the entropy of such a system is constant; we find the same result here by a very different route. The present development thus has yielded some results that are physically plausible, a circumstance providing some assurance that the initial assumptions and mathematical manipulations of our argument retain validity.

2. Moments

Other real quantities related to the $p(\bar{x},t)$ density function that will be of considerable interest to the control system designer are the mean and variance of the state. First, define one element of the mean state vector by

$$h_\ell \overset{\text{def}}{=} \int\!\!\cdots\!\!\int d\bar{x}\; x_\ell p(\bar{x},t) \qquad \ell = 1,2\ldots,n$$

and differentiation with respect to time gives

$$\dot{h}_\ell = \int\!\!\cdots\!\!\int d\bar{x}\; x_\ell \frac{\partial p}{\partial t}$$

$$= -\sum_{k=1}^{n} \int\!\!\cdots\!\!\int dx\; x_\ell \frac{\partial}{\partial x_k}(\alpha_k p) + \frac{1}{2}\sum_{k=1}^{n}\sum_{j=1}^{n} \int\!\!\cdots\!\!\int dx\; x_\ell \frac{\partial^2}{\partial x_k \partial x_j}(\beta_{kj} p)$$

Integration by parts and use of the boundary behavior assumption (b) gives for an integral of the first sum

$$\int..\int d\bar{x}\; x_\ell \frac{\partial}{\partial x_k} (\alpha_k p) \qquad = \qquad - \delta_{\ell k} \int..\int d\bar{x}\; \alpha_\ell p.$$

while the same process shows that all the integrals of the second sum are zero. Recall the definition

$$\bar{\alpha} = F\bar{x} + H\bar{m}$$

and assemble the elements of the mean vector so the set of differential equations for the elements of $\bar{h}$ can be written

$$\dot{\bar{h}} = F\bar{h} + H \int..\int d\bar{x}\; \bar{m}(\bar{x}) p(\bar{x}, t) \tag{7}$$

These may be easy or difficult to evaluate, depending on the form of $\bar{m}(\bar{x})$.

Consider now the variance. Define an element of the covariance matrix of the state by

$$\underline{\gamma}_{ij} \quad = \int..\int d\bar{x}\; \{x_i - h_i\}\; \{x_j - h_j\} p(\bar{x}, t)$$

and differentiate with respect to time to get

$$\dot{\underline{\gamma}}_{ij} = \int..\int d\bar{x}\; \{x_i - h_i\}\; \{x_j - h_j\} \frac{\partial p}{\partial t}$$

the terms containing $\dot{h}_i$, and $\dot{h}_j$ disappearing since

$$\int..\int d\bar{x}\; \{x_i - h_i\}\; \dot{h}_j p = \quad \dot{h}_j \int..\int d\bar{x}\; \{x_i - h_i\} p = 0$$

As before, we substitute for $\frac{\partial p}{\partial t}$ the right-hand side of the Fokker-Planck equation, apply integration by parts to the integrals appearing in the sums, and invoke the boundary behavior assumption (b). The resulting differential equations, if the definitions of $\bar{\alpha}$ and

$$[\beta_{ij}] = GQG^T + HSH^T$$

are recalled, can be assembled and written as

$$[\dot{\gamma}_{ij}] = \dot{\gamma} = F\gamma + \gamma F^T + GQG^T$$

$$+ \left[\int..\int d\bar{x}\ \{\bar{x} - \bar{h}\}m^{-T}p\right]H^T + H\left[\int..\int d\bar{x}\ \bar{m}\{\bar{x} - \bar{h}\}^T p\right]$$

$$+ H\left[\int..\int dx\ S\ p\right]H^T \qquad (8)$$

which shows how the system dynamic effects, F and $H\bar{m}$, and the random effects, Q and S, combined to determine the evolution of the γ covariance matrix.

The same process can be used to obtain differential equations for the third and higher moments. But, instead of writing these out, it will be well to examine the usefulness of the relations already derived. First, note that if the elements of $\bar{m}(\bar{x})$ can be represented by first-degree polynomials in the elements of $\bar{x}$ and $S(\bar{x})$ by second-degree polynomials then Eqs. (7) and (8) for the mean and variance are complete in the sense that the integrals involving $\bar{m}$ and $\bar{S}$ can be written directly as algebraic forms in the elements of $\bar{h}$ and γ, since $p(\bar{x},0)$ is assumed known the initial conditions $\bar{h}(0)$ and $\gamma(0)$ are known and the set of equations for $\bar{h}$ and γ can be integrated directly. This class of problems includes the completely linear case--i.e., $\bar{m}(\bar{x}) = M\bar{x}$ with M a matrix independent of $\bar{x}$, S not dependent on $\bar{x}$--as well as a number of interesting nonlinear examples.

C. DERIVATION OF THE FOKKER-PLANCK EQUATION

The derivation given here of the Fokker-Planck partial differential equation follows the standard methods of proof (see, for example, Wang and Uhlenbeck, "On the Theory of Browian Motion" in Selected Papers on Noise and Stochastic Processes, N. Wax edit., Dover-Publications, New York, 1954). While the problem considered is somewhat more general than those analyzed earlier, this proof is presented primarily for completeness.

We suppose that the system state is described by the n-vector $\bar{x}$, evolution of which is governed by the equations of motion

$$\dot{\bar{x}} = F\bar{x} + G\bar{\omega} + H\bar{y} \qquad 0 \leq t < \infty \qquad (9)$$

with matrices F, G, and H possibly time varying. Let $\bar{\omega}(t)$ be a white random vector with ensemble mean zero and covariance matrix Q,

$$\langle \bar{\omega} \rangle = 0, \ \langle \bar{\omega}(t_1)\omega^{-T}(t_2) \rangle = Q \ \delta \ (t_1 - t_2)$$

Let $p_f(\bar{y}|\bar{x})$ be the conditional probability density of the feedback term $\bar{y}$; write the conditional mean $\bar{m}$ as

$$\bar{m} = \bar{m}(\bar{x}) = \int..\int d\bar{y} \ \bar{y} \ p_f \ (\bar{y}|\bar{x})$$

and the ensemble conditional covariance as

$$\langle \{\bar{y}(t_1) - \bar{m}_1\} \ \{y(t_2) - \bar{m}_2\}^T | \ \bar{x}(t_1), \ \bar{x} \ (t_2) \rangle = S \ \delta \ (t_1 - t_2)$$

$$= S \ (\bar{x}) \ \ \delta \ (t_1 - t_2)$$

$$= \delta \ (t_1 - t_2) \int..\int \bar{dy}\{\bar{y} - \bar{m}\}\{\bar{y} - \bar{m}\}^T p_f(\bar{y}|\bar{x})$$

so that $\bar{y}$ is probabilistically dependent on $\bar{x}$ and is white in time. Take $\bar{y}$ and $\bar{\omega}$ to be statistically independent, write p(x,t) for the ensemble probability density of the system state at any time t, and take $p(\bar{x},0)$ to be given.

Because the system state at any time is sufficient to determine later states given the inputs over the intervening time interval, and because the "inputs" $\bar{\omega}$ and $\bar{y}$ here are white in time, the evolution of the state $\bar{x}(t)$ is a Markov process. Therefore we can define the transition probability function

$$q(\bar{u},t|\bar{v},t_1)d\bar{u} = \text{Prob}(u_1 < x_1(t) < u_1 + du_1, \ldots,$$

$$u_n \leq x_n(t) < u_n + du_n \text{ given that } \bar{x}(t_1) = \bar{v})$$

$$\int\ldots\int d\bar{x} q(\bar{x},t|\bar{v},t_1) = 1 \tag{10}$$

and we have the basic relation

$$p(\bar{x},t) = \int\ldots\int d\bar{v}\ p(v,t_1)\ q(\bar{x},t|\bar{v},t_1)$$

from which we get

$$\frac{1}{\tau}\{p(\bar{x},t+\tau) - p(\bar{x},t)\} = \frac{1}{\tau}\int\ldots\int d\bar{v}\ p(\bar{v},t)q(\bar{x},t+\tau|\bar{v},t) - \frac{1}{\tau}\,p(\bar{x},t)$$

Now take $f(\bar{x})$ to be arbitrarily chosen real function that is analytic in $\bar{x}$ and integrable over the state space; also require that $f(\bar{x})$ and all its partial derivatives vanish as $\bar{x} \to \infty$. Multiply the above equation by $f(\bar{x})$, integrate over the state space, and let τ become small so that

$$\lim_{\tau \to 0} \int\ldots\int d\bar{x}\ f(\bar{x})\ \frac{p(\bar{x},t+\tau) - p(\bar{x},t)}{\tau} = \int\ldots\int d\bar{x}\ f(\bar{x})\ \frac{\partial p}{\partial t}$$

$$= \lim_{\tau \to 0} \frac{1}{\tau}\left\{\int\ldots\int d\bar{x}\int\ldots\int d\bar{v}\ f(\bar{x})p(\bar{v},t)q(\bar{x},t+\tau|\bar{v},t)\right. \tag{11}$$

$$\left. - \int\ldots\int d\bar{x}\ f(\bar{x})p(\bar{x},t)\right\}$$

Expand $f(\bar{x})$ in Taylor Series, with $\bar{z} = \bar{x} - \bar{v}$,

$$f(\bar{x}) = f(\bar{v} + \{\bar{x} - \bar{v}\}) = f(\bar{v} + \bar{z}) \tag{12}$$

$$= f(\bar{v}) + \sum_{k=1}^{n} z_k \frac{\partial f(\bar{v})}{\partial v_k} + \frac{1}{2}\sum_{k=1}^{n}\sum_{j=1}^{n} z_k z_j \frac{\partial^2 f(\bar{v})}{\partial v_k \partial v_j} + \ldots$$

and the right-hand side of the above equation becomes

$$\lim_{\tau \to 0} \frac{1}{\tau} \left\{ \int\!\!\cdots\!\!\int d\bar{v}\ f(\bar{v}) p(\bar{v},t) \int\!\!\cdots\!\!\int d\bar{x}\ q(x, t+\tau | \bar{v}, t) \right.$$

$$+ \sum_{k=1}^{n} \int\!\!\cdots\!\!\int d\bar{v}\ p(\bar{v},t) \frac{\partial f(\bar{v})}{\partial v_k} \int\!\!\cdots\!\!\int d\bar{x}\ z_k\ q(\bar{x}, t+\tau | \bar{v}, t)$$

$$+ \frac{1}{2} \sum_{k=1}^{n} \sum_{j=1}^{n} \int\!\!\cdots\!\!\int d\bar{v}\ p(\bar{v},t) \frac{\partial^2 f(\bar{v})}{\partial v_k \partial v_j} \int\!\!\cdots\!\!\int d\bar{x}\ z_k z_j\ q(\bar{x}, t+\tau | \bar{v}, t)$$

$$\left. + \ \ldots\ . - \int\!\!\cdots\!\!\int d\bar{x}\ f(\bar{x})\ p(\bar{x}\ t) \right\}$$

Because of the normalization of $q(\bar{x}\ t_1 | \bar{v}\ t)$ the first and last integrals here subtract to zero.

Define

$$\alpha_k(\bar{v}) = \lim_{\tau \to 0} \frac{1}{\tau} \int\!\!\cdots\!\!\int d\bar{z}\ z_k\ q(\bar{z} + \bar{v},\ t+\tau | \bar{v}\ t)$$

$$\beta_{kj}(v) = \lim_{\tau \to 0} \frac{1}{\tau} \int\!\!\cdots\!\!\int d\bar{z}\ z_k z_j\ q(\bar{z} + \bar{v},\ t+\tau | \bar{v}, t)$$

and the right-hand expression is

$$\sum_{k=1}^{n} \int\!\!\cdots\!\!\int d\bar{v}\ p(\bar{v},t) \alpha_k(\bar{v}) \frac{\partial f(\bar{v})}{\partial v_k}$$

$$+ \frac{1}{2} \sum_{k=1}^{n} \sum_{j=1}^{n} \int\!\!\cdots\!\!\int d\bar{v}\ p(\bar{v},t)\ \beta_{kj}(\bar{v}) \frac{\partial^2 f(\bar{v})}{\partial v_k \partial v_j} + \ \ldots$$

Integrate the integrals in the first sum by parts and recall the boundary assumptions on $f(\bar{x})$ to get, for example,

$$\int\!\!\cdots\!\!\int d\bar{v}\ p(\bar{v},t) \alpha_k(\bar{v}) \frac{\partial f(\bar{v})}{\partial v_k} = - \int\!\!\cdots\!\!\int d\bar{v}\ f(\bar{v}) \frac{\partial}{\partial v_k} (\alpha_k p)$$

Similarly, the typical integral of the second summation can be written

$$\int..\int dv\ p(\bar{v},t)\beta_{kj}(\bar{v})\ \frac{\partial^2 f(\bar{v})}{\partial v_k \partial v_j} - \int..\int d\bar{v}\ f(\bar{v})\ \frac{\partial^2}{\partial v_k \partial v_j}\ (\beta_{kj}p)$$

The same type of operation can be applied to all the other terms in the Eq. (12), and so Eq. (11) becomes when all terms are collected on the left-hand side,

$$\int..\int d\bar{x}\ f(\bar{x})\ \left\{\frac{\partial p}{\partial t} + \sum_{k=1}^{n}\ \frac{\partial}{\partial x_k}\ (\alpha_k p)\right.$$

$$\left. - \frac{1}{2}\ \sum_{k=1}^{n}\ \sum_{j=1}^{n}\ \frac{\partial^2}{\partial x_k \partial x_j}\ (\beta_{kj}p) + \ldots\right\} = 0$$

The function $f(\bar{x})$ is arbitrary except for some mild smoothness and boundary conditions, and so we find that $p(\bar{x},t)$, must satisfy

$$\frac{\partial p}{\partial t} = -\sum_{k=1}^{n}\ \frac{\partial}{\partial x_k}\ (\alpha_k p) + \frac{1}{2}\ \sum_{k=1}^{n}\ \sum_{j=1}^{n}\ \frac{\partial^2}{\partial x_k \partial x_j}\ (\beta_{kj}p) + \ldots \qquad (13)$$

The only remaining steps are to calculate the α's and β's, and to show that we need only the first- and second-order terms in Eq. (12). Let $X(t)$ be the "fundamental matrix" of Eq. (9) satisfying

$$\dot{X} = FX\ ,\quad X(0) = I$$

and if $\bar{x}(t) = \bar{v}$ we have

$$\bar{x}(t + \tau) = X(t + \tau)X^{-1}(t)\bar{v} + X(t + \tau)\int_t^{t+\tau} dt_1 X^{-1}(t_1)G(t_1)\bar{\omega}(t_1)$$

$$+ X(t + \tau)\int_t^{t+\tau} dt_1 X^{-1}(t_1)H(t_1)\bar{y}(t_1)$$

The step in state space from time t to $t + \tau$ was designated $z = x(t + \tau) - v$; further, for τ small the differential equation for X shows that we can write

$$X(t + \tau) \simeq X(t) + \tau F X(t)$$

With these substitutions the equation for $\bar{x}(t + \tau)$ becomes

$$z \simeq \tau F\bar{v} + [I + \tau F]X(t) \int_t^{t+\tau} dt_1 X^{-1}(t_1)G(t_1)\bar{\omega}(t_1)$$

$$+ [I + \tau F]X(t) \int_t^{t+\tau} dt_1 X^{-1}(t_1)H(t_1)\bar{y}(t_1)$$

In computing the α's, we take $\bar{v}$ given as the state at time t and calculate the average $\bar{z}$ over all transitions. This is the same as averaging $\bar{z}$ over the ensemble of $\bar{\omega}$ and $\bar{y}$ with the condition that $\bar{v}$ is given; therefore, since $\langle\bar{\omega}\rangle = 0$,

$$\langle\bar{z}\,|\,\bar{x} = \bar{v}\rangle \simeq \tau F\bar{v} + [I + \tau F]X(t) \int_t^{t+\tau} dt_1 X^{-1}(t_1)H(t_1)\bar{m}_1$$

and we have

$$\bar{\alpha}(\bar{v}) = \lim_{\tau \to 0} \frac{1}{\tau} \langle\bar{z}\,|\,\bar{x} = \bar{v}\rangle = F\bar{v} + H\bar{m}(\bar{v})$$

Similarly, for the β's we compute the conditional ensemble average

$$\langle\bar{z}\bar{z}^T\,|\,\bar{x} = \bar{v}\rangle = X(t) \int_t^{t+\tau} dt_1 \int_t^{t+\tau} dt_2 X^{-1}(t_1)G(t_1) \langle\omega^{-T}(t_1)\omega^{-T}(t_2)\rangle$$

$$G^T(t_2)X^T(t)$$

$$+X(t) \int_t^{t+\tau} dt_1 \int_t^{t+\tau} dt_2 X^{-1}(t_1)H(t_1) \langle\bar{y}(t_1)y^{-T}(t_2)\,|\,\bar{v}\rangle H^T(t_2)X^{-T}(t_2)X^T(t)$$

$$+ \text{terms of order } \tau^2$$

which as a result of the whiteness assumptions for $\bar{\omega}$ and $\bar{y}$ becomes

$$\langle \bar{z}\bar{z}^T | v \rangle = X(t) \int_t^{t+\tau} dt_1 X^{-1}(t_1) G(t_1) Q(t_1) G^T(t_1) X^{-T}(t_1) X^T(t)$$

$$+ X(t) \int_t^{t+\tau} dt_1 X^{-1}(t_1) H(t_1) S(\bar{v}) H^T(t_1) X^{-T}(t_1) X^T(t)$$

$$+ \text{terms of order } \tau^2$$

Thus we get the matrix $[\beta_{kj}]$

$$[\beta_{kj}] = \lim_{\tau \to 0} \frac{1}{\tau} \langle \bar{z}\bar{z}^T | \bar{x} = \bar{v} \rangle$$

$$= GQG^T + HS(\bar{v})H^T$$

Finally, all terms in Eq. (12) of order higher than second will involve moments of $\bar{z}$ of order higher than two. In the limit as the time step τ goes to zero the effect of these higher moments in the partial differential equation for $p(\bar{x},t)$ will be negligible, and so we need only terms to second-order. This completes the argument.

APPENDIX F

APPROXIMATE DESIGN OF A FIXED CONTROLLER SYSTEM WITH DIGITAL FEEDBACK

The example problem treated in the course of the study is formulated in Sec. IV-D of Vol. I, and the main results are given in the form of a design chart. The purpose of the present Appendix is to substantiate these results.

A. PROBLEM FORMULATION AND NOMENCLATURE

It is desired to find the optimum design parameters of the digital position control servo shown in Fig. F-1.

Nomenclature

The following nomenclature is used throughout the appendix.

$[-S_B, S]$ = quantizer range and input command range

N = number of quantization levels

L = number of bits per message

δt = bit time

ΔT = sampling period and message time

H = quantization increment size

p = bit error probability

$p(x)$ = output probability

$\vartheta(t)$ = impulse response of the continuous and linear system

$\underline{x}$ = state of the system; the state has 3 components

x_1 = output position

x_2 = output velocity

x_3 = decoder output

u = control signal

$r(t)$ = command input

R = constant value of command input

v = noise affecting the channel

J = operating cost

s = Laplace transform operator

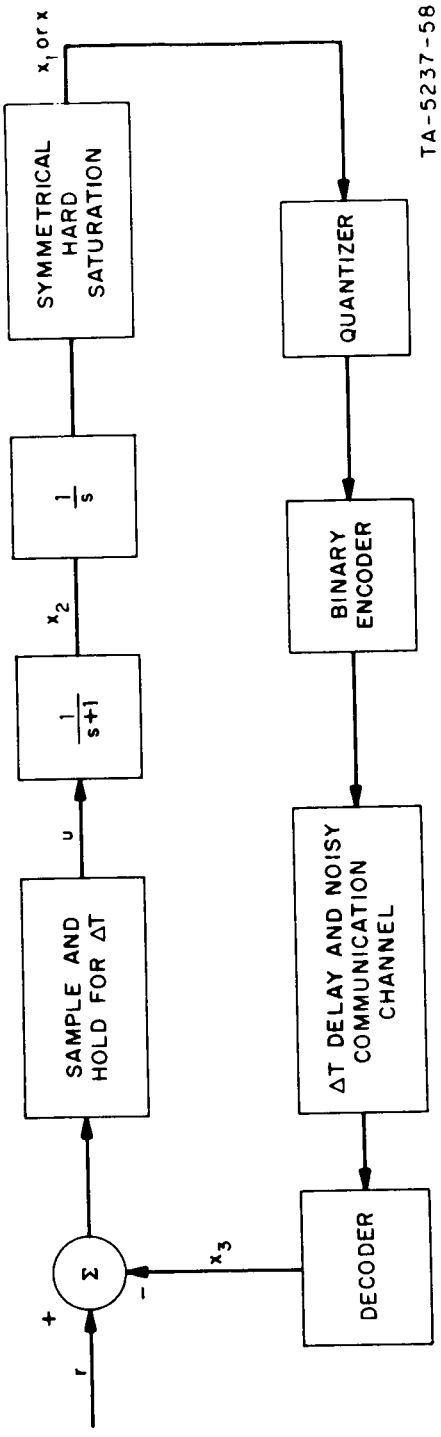

FIG. F-1 SAMPLED DATA FEEDBACK CONTROL SYSTEM WITH SIGNAL-DEPENDENT FEEDBACK NOISE

B. SYSTEM DESCRIPTION

As shown in Fig. F-1 the plant is characterized by the transfer function

$$\frac{X(s)}{u(s)} = \frac{1}{s(s+1)} \tag{1}$$

Due to the hold action ahead of the plant, we find

$$\ddot{x}(t) + \dot{x}(t) = u \tag{2}$$

where u is a constant in the interval ΔT. Transforming we find

$$X(s) = \frac{x(0)\, s(s+1) + \dot{x}(0)\, s + u}{s^2(s+1)}$$

$$= \frac{x(0)}{s} + \dot{x}(0)\left[\frac{1}{s} - \frac{1}{s+1}\right] + u\left[\frac{1}{s^2} - \frac{1}{s} + \frac{1}{s+1}\right] \tag{3}$$

or

$$x(t) = x(0) + \dot{x}(0)\,[1 - e^{-t}] + u\,[t - (1-e^{-t})] \tag{4}$$

and

$$\dot{x}(t) = \dot{x}(0)\, e^{-t} + u[1 - e^{-t}]. \tag{5}$$

If now we consider the state vector $\underline{x}(t)$ we find

$$\underline{x}(t) = \begin{bmatrix} x(t) \\ \dot{x}(t) \end{bmatrix} = \begin{bmatrix} 1 & 1-e^{-t} \\ 0 & e^{-t} \end{bmatrix} \underline{x}(0) + u \begin{bmatrix} t - (1-e^{-t}) \\ 1 - e^{-t} \end{bmatrix} \tag{6}$$

Turning now to the sample-time ΔT we find

$$\underline{x}(\Delta T) = \begin{bmatrix} 1 & V \\ 0 & U \end{bmatrix} \underline{x}(0) + u(0) \begin{bmatrix} \Delta T - V \\ V \end{bmatrix} \tag{7}$$

where, for convenience, we have set

$$U = e^{-\Delta T} \qquad a$$

$$V = 1 - e^{-\Delta T} \qquad b \tag{8}$$

Ignoring all the feedback complications except the delay we can write

$$\underline{x}\ (n+1) = \begin{bmatrix} 1 & V \\ 0 & U \end{bmatrix} \underline{x}\ (n) + u(n) \begin{bmatrix} \Delta T - V \\ V \end{bmatrix} \qquad a \qquad (9)$$

$$u(n) = R(n) - x_1(n-1) \qquad b$$

where these equations are valid only at the sampling instants when $t = n\Delta T$.

Substituting (9b) in (9a),

$$\underline{x}\ (n+1) = \begin{bmatrix} 1 & V \\ 0 & U \end{bmatrix} \underline{x}\ (n) + R(n) \begin{bmatrix} \Delta T - V \\ V \end{bmatrix} + \begin{bmatrix} V - \Delta T & 0 \\ -V & 0 \end{bmatrix} \underline{x}\ (n-1) \qquad (10)$$

Rewriting (10) as separate equations, calling the final $x(n-1)$, $y(n-1)$, to indicate the effect of the communications channel, $\dot{x}(n)$ can be eliminated to give

$$x(n+3) = (1+U)\ x\ (n+2) - Ux(n+1) + (V-\Delta T)\ y(n+1) + (U\Delta R - V)\ y(n) + (\Delta T - V)$$

$$R(n+1) + (V - U\Delta T)\ R(n) \qquad (11)$$

where use has been made of the equality

$$UV + V^2 = V(U+V) = V \qquad (12)$$

A block diagram of (11) is shown in Fig. F-2 where

$$A_1 = 1 + U \qquad a$$

$$A_2 = -U \qquad b$$

$$A_3 = V - \Delta T \qquad c \qquad (13)$$

$$A_4 = U\Delta T - V \qquad d$$

1. Saturation, Quantizing, Coding, and Error

As shown in Fig. F-2, the present output, $x(n)$, is hard limited to an arbitrary level which is actually set to ± 1.5 to correspond to an expected range of inputs from -1 to $+1$.

Since the communication channel is binary it can handle only discrete messages. Since no error correction or detection coding was

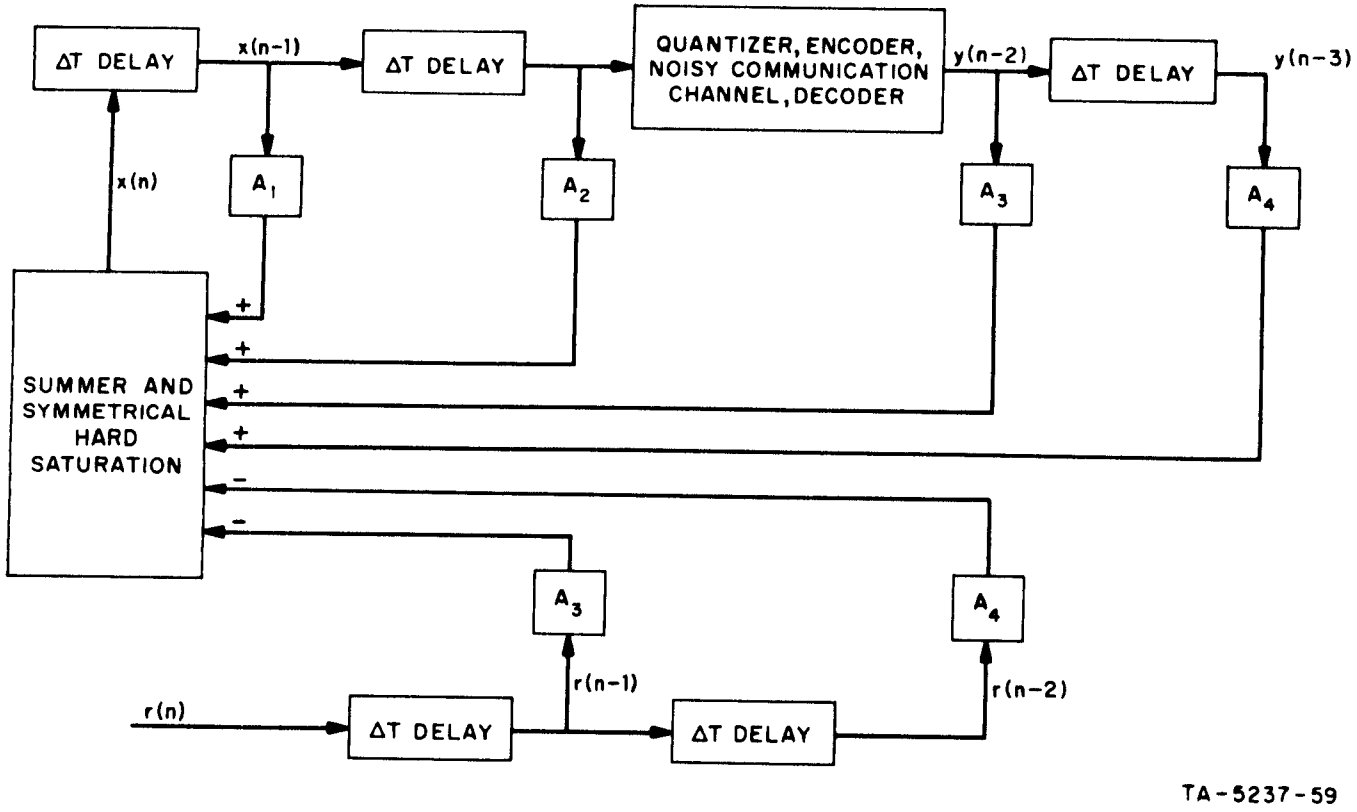

FIG. F-2 EQUIVALENT BLOCK DIAGRAM OF SAMPLED DATA FEEDBACK CONTROL SYSTEM WITH SIGNAL-DEPENDENT FEEDBACK NOISE

considered, the number of quantizing levels is a power of 2, actually 2^L where L is the number of bits per message. Certainly one level should be zero and therefore it is impossible for the remaining levels to be symmetrically located about zero. Hence there are 2^{L-1} positive levels, a zero level, and $2^{L-1}-1$ negative levels. As an example for l=3, when

$-\infty < x < -.9375$,	the level is -1.125
$-.9375 < x < -.5625$	- .75
$-.5625 < x < -.1875$	- .375
$-.1875 < x < .1875$	0
$.1875 < x < .5625$	+ .375
$.5625 < x < .9375$	.75
$.9375 < x < 1.2125$	1.125
$1.2125 < x < \infty$	1.500

The possible effects of this asymmetry obviously decrease with increasing L.

The levels have been coded with the most negative represented by L zeroes increasing in regular binary to the most positive represented by L ones. The channel is considered symmetrical with a bit error probability of a 1 becoming a 0 or a 0 becoming a 1 equal to p. Errors are assumed to occur independently in each bit so that for L bits, the probability of a single error is $p(1-p)^{L-1}$ of two errors $p^2(1-p)^{L-2}$, etc. Note that the effect of a bit error depends on the transmitted level and on the significance of the particular bit. The size of errors in level due a single bit error vary from $1.5/2^{L-1}$ to 1.5.

The operation of decoding consists here merely of converting the received binary word into the equivalent quantized level.

2. The Channel

It is assumed that the digital channel is of the frequency shift keying variety (F.S.K.) for which the bit error probability is

$$p = 1/2\ e^{-\alpha\delta t} \tag{14}$$

C. SAMPLING EFFECTS

1. General

A program was prepared to simulate the system of Fig. F-1 without quantization, saturation, or feedback noise to determine the integral squared error of the response following a feedback signal offset. The appropriate difference equation is

$$x(n+3) = A_1 x(n+2) + (A_2 + A_3)\, x(n+1) + A_4 x(n) \qquad (15)$$

where

$$\begin{aligned} A_1 &= 1 + U && \text{a} \\ A_2 + A_3 &= 1 + V - \Delta T && \text{b} \\ A_4 &= U\Delta T - V && \text{c} \end{aligned} \qquad (16)$$

and x(n) is the output at time n.

The integral square error was approximated by the sum of x(n) squared at the sampling instants over a ten-second simulated time interval:

$$ISE = \sum_n x^2(n)\, \Delta T \qquad (17)$$

The continuous case (equivalent to $\Delta T = 0$) was included for comparison.

A program was also prepared to determine the best second order polynomial fit to the integral square error for the position and velocity displacement cases.

2. Results

The results of these simulations are presented in the following figures:

Figure F-3: $1/\Delta T$ Feedback Displacement

Output vs. Time

a. Continuous (Amplitude = 1)

b. $\Delta T = 0.1$ (Amplitude = 10)

c. $\Delta T = 0.2$ (Amplitude = 5)

d. $\Delta T = 0.4$ (Amplitude = 2.5)

Figure F-4: Unit Feedback Displacement

Log ISE vs. log ΔT

a. Simulation

b. ΔT^2 · (ISE for continuous case)

Based on computer runs not plotted here, the system goes unstable for ΔT larger than some value between 0.7 and 0.8 seconds.

The key result is Figure F-4 which shows that the integral square error is well approximated by $M\Delta T^2$, M = constant, for $0 \leq \Delta T \leq 0.2$.

D. CODING CONSIDERATIONS FOR SAMPLED DATA CONTROL SYSTEMS WITH A BINARY FEEDBACK COMMUNICATION CHANNEL

1. General

In the usual binary communication channel the available alphabet is finite and is equal to or less than 2^K where K is the number of bits per message.* There are two general types of redundant codes: systematic and non-systematic.

We will restrict our considerations to symmetric channels where the probabilities of receiving a one when a zero was sent and of receiving a zero when a one was sent are equal. We will also assume that the bit errors occur independently and that the receiver makes a decision on each bit.

---

*There are, of course, codes where the number of bits varies from word to word. The present section is restricted to uniform sampling and a fixed number of bits per word.

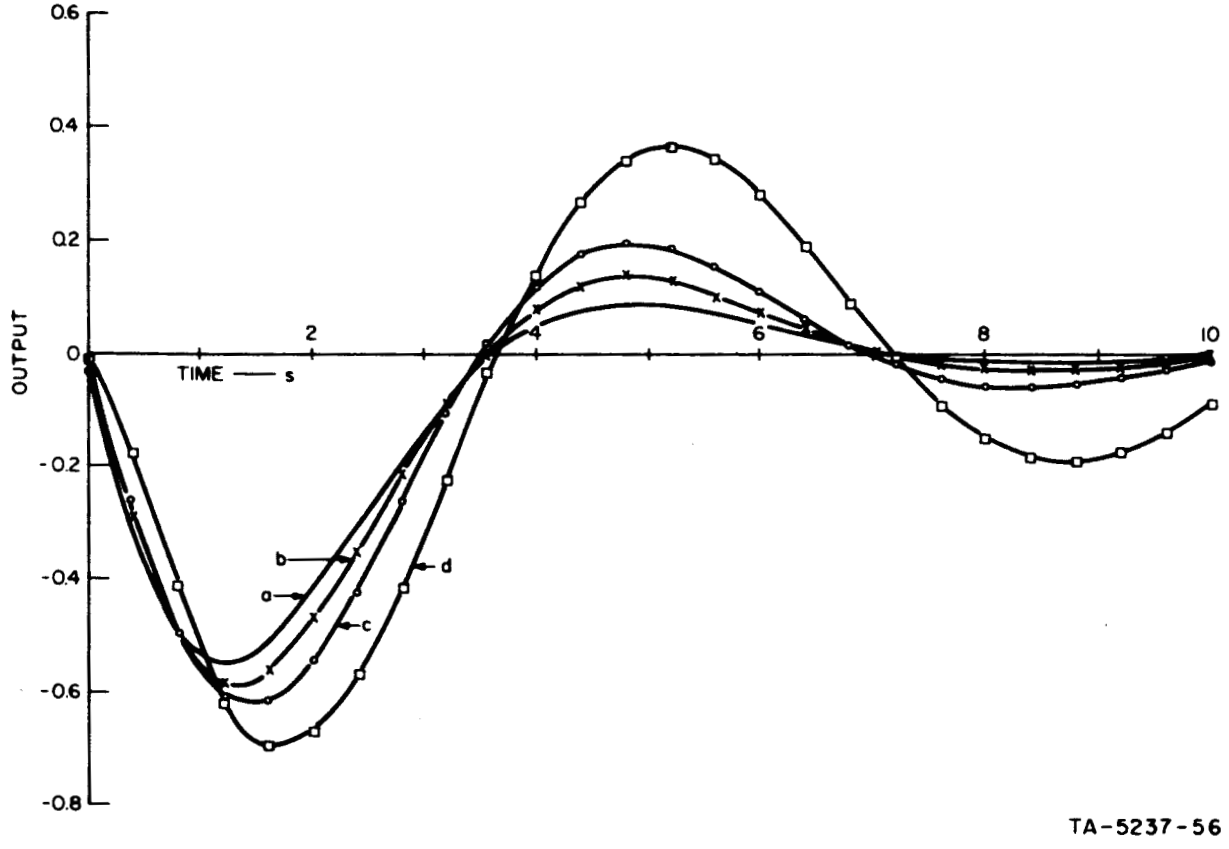

FIG. F-3 OUTPUT FOLLOWING A FEEDBACK DISPLACEMENT

(a) Continuous System Amplitude 1
(b) Sampled at $\Delta T = 0.1$ Amplitude 10
(c) Sampled at $\Delta T = 0.2$ Amplitude 5
(d) Sampled at $\Delta T = 0.4$ Amplitude 2.5

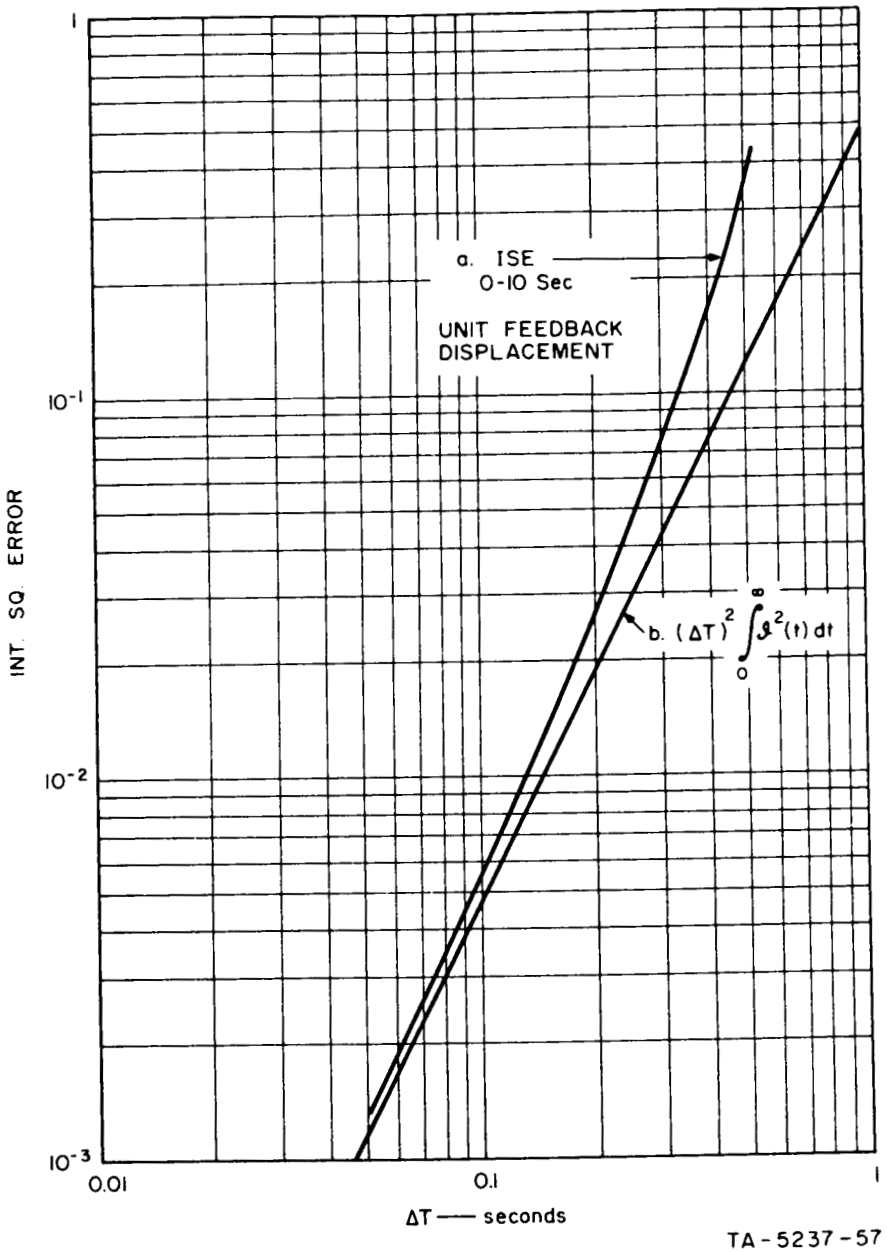

FIG. F-4 Log ISE vs. Log ΔT
(a) Simulation
(b) ΔT^2 ISE for Continuous Case

2. Systematic Codes

In a systematic code the size of the alphabet is 2^K where $K \leq L$ is the number of information bits and where the K - L redundant bits are determined by parity check equations.

The power of such codes is measured by the minimum distance between any two code words of the alphabet. The redundancy may be used either for error correction or error detection or both. Protection is thus provided against some numbers of errors per word but never against all possible errors.

a. Error Correction

The use of redundancy to provide error correction is much less attractive than is popularly believed. Under the usual constraints of fixed power and data rate, the addition of redundancy must reduce the time per bit and increase the bit error probability accordingly. Only when the bit error probability is very low without the extra bits required for error correction does the error correction capability increase more rapidly than the number of errors per message which increases both because of less time per bit and because of more bits per word. Since we are primarily interested in relatively poor channels, error correction will not be further considered.

b. Error Detection

The use of available redundancy to detect the occurrence of errors is generally an effective technique but raises the obvious question in the feedback control context of what action is to be taken when errors are detected. The possibilitites include:

i setting the feedback to zero

ii setting the feedback equal to the command

iii using the previous value (which must then have been saved)

iv using a function of the previous or several previous values.

Without any implication that such procedures would not be useful, they have not been explored further in the present project.

3. Non-Systematic Codes

In non-systematic codes information bits as such cannot be identified. Rather the code words, however structured, are assigned to the various messages at the user's convenience. Examples include the Gray codes where the code words assigned to adjacent levels differ by one bit and the "Snake-in-the-Box" codes* which are extended Gray codes in that, up to some difference in levels, the words assigned differ in the same number of bits as the difference in levels to which they are assigned. These are primarily error detection codes and possess no specifically desirable properties for the present purpose. Since their use requires a decision rule as outlined in II-B above, they are not further considered here.

4. Other Techniques

For any reasonable channel error probability, reception of a word with no errors is more probable than with single errors which in turn is more probable than with double errors, etc. If the system is to be operated as a regulator so that the commanded output level is fixed, any code can be rearranged to make the effect of single bit errors less serious by assigning to those levels near the commanded level code words which differ in only a single bit from the word assigned to the commanded level. This technique is obviously applicable only when the commanded level is fixed and the output does in fact stay near that level.

Probably the most powerful technique for controlling error effects in a feedback system is to devote extra time and/or power to the more significant bits of the message. Additional terminal complexity is,

---

*Singleton, R. C., "Generalized Snake in the Box Codes," to be published in the IEEE Transactions on Electronic Computers.

of course, required. Such systems have been discussed in the literature* and may actually have been implemented but will not be further considered here.

5. Conclusions

Primarily for simplicity, the present work has considered only the use of non-redundant binary codes in which the all zeroes word has been assigned to the most negative level, the word having zero in its most significant bit and ones elsewhere to the zero level, and the all ones word to the most positive level.

E. DEADBAND COST

If the system is operating in the transient mode, a cost J_2', referred to as the deadband cost, is incurred. Even in the absence of communication channel errors, the output x does not exactly equal the command input r in general because of the quantization deadband. In general, this error is not constant, i.e., there is limit cycle motion x(t) contained roughly within this deadband. In order to evaluate approximately the degrading effects of these errors, it is assumed that the output probability p(x) is uniform within the deadband H. Consequently,

$$J_2' = \int_{-\frac{H}{2}}^{\frac{H}{2}} x^2 \, p(x) \, dx = \frac{H^2}{12} \tag{18}$$

F. TRANSIENT COST

If the system is operating in the transient mode, an erroneous signal w is occasionally received at the decoder output. This signal,

---

*Bedrosian, Edward, "Weighted PCM," IRE Trans., Vol IT-4, #1, March 1958, pp. 45-49, and Bellman, R. and Kalaba, R., "On Weighted PCM and Mean-Square Deviation," IRE Trans., Vol. IT-4, #1, March 1958, pp. 58-59.

which is applied during the interval ΔT, sets up a transient the degrading effect of which is measured by the integral squared error (ISE) discussed in Section C, namely

$$ISE = w^2 \int_0^\infty \vartheta^2 (t, \Delta T)\, dt \tag{19}$$

where $\vartheta (t, \Delta T)$ is the response of the linearized closed-loop system to a unit perturbation w applied during ΔT. It is seen from Fig. F-4 that the response $\vartheta (t, \Delta T)$ is well approximated by

$$\vartheta (t, \Delta T) = \Delta T\, \vartheta (t, 0) \tag{20}$$

where $\vartheta (t, 0)$ is defined as the unit impulse response of the continuous system containing no communication channel delay ΔT.

With the response $\vartheta (t, \Delta T)$, there is associated the integral squared error $M (\Delta T)$

$$M(\Delta T) = \int_0^\infty \vartheta^2 (t, \Delta T)\, dt$$

It is seen from Fig. F-4 that

$$M(\Delta T) \cong (\Delta T)^2 M(0)$$

where M(0) is the ISE of the continuous system perturbed by a unit impulse.

This useful, but certainly not necessary simplification, is retained in the remainder of this appendix.

1. The Transient Cost J_2''

If the bit error probability is sufficiently small, the transient generated by a first perturbation w will have subsided on the average, by the time a second perturbation w occurs. This is a valid

assumption in the transient mode of operation. The resulting per unit time cost J_2'' is then defined by

$$J_2'' \triangleq \int_W M(\Delta T)\, w^2\, p(w)\, dw = M(\Delta T) \int_W w^2 p(w) dw \qquad (21)$$

2. Calculation of $\int_W w^2 p(w) dw$

Once a particular code has been selected, the expression $\int_W w^2 p(w) dw$ can be calculated in a fairly straightforward fashion in terms of the bit error probability p, as will be illustrated below for a two-bit binary code.

Message Sent	Message Received with Prob. 1-p	Message Received with Prob. p(1-p)	Message Received with Prob. p^2
0 0	0 0	01 and 10	1 1
0 1	0 1	00 and 11	1 0
1 0	1 0	11 and 00	0 1
1 1	1 1	10 and 01	0 0

If p is sufficiently small, messages with a single bit error are received with probability $p(1-p) \cong p$ and messages with double bit errors can be neglected.

The perturbations w resulting from single bit errors are shown in terms of the message sent.

Message Sent	Message Received with Prob. p	Perturbation w
0 0	01 and 10	H, 2H
0 1	00 and 11	-H, 2H
1 0	11 and 00	H; -2H
1 1	10 and 01	-H; -2H

If all messages sent are equally likely (which is a convenient, but not necessary assumption,) then

$$\int_w w^2 p(w)dw \cong H^2 p(1+4+1+4+1+4+1+4) = 20H^2 p \qquad (22)$$

This procedure is easily generalized as follows. Let the transmitted message be

$$x = H(S_B + \sum_{j=0}^{L-1} a_j 2^j) \qquad (23)$$

with a_j equal to zero or one.

Then, for each message sent, perturbations of magnitude H_2^j (j = 0, ..., L-1) occur with equal probability. If in addition the transmitted messages are all equally likely, then

$$\int_w w^2 p(w)dw \cong \frac{p2^L H^2}{L-1} \sum_{j=0}^{L-1} 2^{2j} \qquad (24)$$

$$\cong p \frac{4S^2}{2^L(L-1)} \sum_{j=0}^{L-1} 2^{2j} \triangleq pC(L) \qquad (25)$$

The function C(L) is tabulated below for S=2 in terms of L.

L	C(L)
2	20
3	21
4	21.3
5	21.4
6	21.4
7	21.4

It is seen that for $L \geq 5$, the function C(L) is constant.

For p sufficiently small, the transient cost J_2'' is thus approximately

$$J_2'' = M(\Delta T)\ p\ C(L) \tag{26}$$

G. MONTE CARLO RESULTS

1. General

The system was simulated on the computer by the following equation:

$$X[I] = A_1X[I-1] + A_2X[I-2] + A_3Y[I-2] + A_4[I-3] + A_5 \cdot R \tag{27}$$

where X[I] is the output, R is the command (step function) and Y[I] is the feedback signal after quantization and a probability trial to determine a signal dependent error value corresponding to a specified probability of error. The constants are

$$A_1 = 1 + U \tag{28a}$$

$$A_2 = -U \tag{28b}$$

$$A_3 = V - DT \tag{28c}$$

$$A_4 = U \cdot DT - V \tag{28d}$$

$$A_5 = V \cdot DT \tag{28e}$$

where

$$U = 1 - V = e^{-DT} \tag{28f}$$

and DT is the sampling time.

Initial conditions were set to suppress any initial transient; i.e., the system was assumed at rest for $I < 0$ with $X[I] = R$.

The system operated for approximately 10 seconds (simulated time) and the integral square error accumulated as

$$ISE = \sum_{I} (X[I] - R)^2 \cdot DT \tag{29}$$

The values in the following are the integral square error per second

$$J = Cost = \frac{ISE}{I_{max} \cdot DT} \tag{30}$$

The probability distribution of the quantized output was also accumulated and the mean calculated at the end of the run.

2. Transient Mode

Results for R = 0 and operation in the transient mode are given in Table I. Note that the simulation cost does not include dead-band cost. The cost is added to the simulation results as

$$J_{\delta} = \frac{.750}{N^2} \tag{31}$$

where N is the number of quantization levels.

3. Steady State Mode

Results for R = 0.868 and operation in the steady-state mode are presented in Table II. The command was chosen to simulate the costs associated with a uniform probability density for R. In general, after 10 seconds, the output was still cycling and the amplitude of the last full half-cycle is included in Table II both in absolute units and as a percentage of the quantization level.

TABLE I

Transient Mode - R = 0

Number of Bits	Bit Error Probability	Sampling Time	Mean	Simulated Cost	Deadband Cost	Total Cost
2	0.00004	0.3	0.00	0.0469	0.0469	0.0469
3	0.00095	0.3	0.00	0.0117	0.0117	0.0117
4	0.00430	0.3	0.00	0.0061	0.0028	0.0089
5	0.01115	0.3	0.01	0.0085	0.0007	0.0092
3	0.00011	0.4	0.00	0.0000	0.0117	0.0117
4	0.00095	0.4	0.00	0.0000	0.0028	0.0028
5	0.00315	0.4	0.00	0.0048	0.0007	0.0055

TABLE II

Steady State Mode - R = 0.868

Number of Bits	Bit Error Probability	Sampling Time	Mean	Cost	Final 1/2-Cycle Amplitude	
					Absolute	% of Interval
2	0.01	0.1252	0.871	0.2760	0.287	38
3	0.01	0.1878	0.885	0.0448	0.145	39
4	0.01	0.2504	0.877	0.0214	0.116	62
5	0.01	0.3130	0.791	0.0199	0.240	256
2	0.02	0.1016	0.878	0.3285	0.159	21
3	0.02	0.1524	0.897	0.0565	0.101	27
4	0.02	0.1524	0.911	0.0681	0.078	42
5	0.02	0.2540	0.820	0.0615	0.698	744
2	0.05	0.0725	0.896	0.5500	0.163	22
3	0.05	0.1088	0.931	0.1332	0.079	21
4	0.05	0.145	0.951	0.1438	0.265	141
5	0.05	0.1813	0.941	0.1961	0.117	125
2	0.10	0.0509	0.977	1.0100	0.105	14
3	0.10	0.0764	0.055	0.5950	0.345	92
4	0.10	0.1018	1.034	0.368	0.045	24
5	0.10	0.1273	1.110	0.644	0.146	156